DPTO. TEORÍA DE LA SEÑAL Y COMUNICACIONES

INGENIERÍA TÉCNICA DE TELECOMUNICACIÓN: SONIDO E IMAGEN

PROYECTO FIN DE CARRERA

SEGMENTACIÓN DEL MOVIMIENTO EN SECUENCIAS DE VÍDEO Y SU APLICACIÓN A LA CODIFICACIÓN PERCEPTUAL DE VÍDEO

Autor: Ana Belén Mejía Ocaña

Tutor: Sergio Sanz Rodríguez-Escalona

Leganés, 20 de julio de 2010

Título:

Autor:

Director:

EL TRIBUNAL

Presidente:

Vocal:

Secretario:

Realizado el acto de defensa y lectura del Proyecto Fin de Carrera el día ____ de _____ de 20___ en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

VOCAL

SECRETARIO PRESIDENTE

Agradecimientos

Este documento supone el final de un largo camino y el cese de una etapa de especial relevancia en mi vida. Por ello, he de echar la vista atrás para recordar todo el esfuerzo, sacrificio, interés, ilusión y responsabilidad puestos con el objetivo de llegar a este momento y conseguir enorgullecer a todos aquellos que siempre han confiado en mí.

A lo largo de este tiempo, he contado con el apoyo incondicional de mi familia que he de agradecer enormemente. Gracias por toda la preocupación y dedicación mostrada en forma de llamadas, charlas y, avisos que, en ocasiones, he repudiado y que sin embargo, me han hecho reflexionar y continuar hacia delante.

Gracias a aquellas personas que llevan a mi lado durante muchos años, en los buenos y malos momentos, y que por supuesto, han sido un gran apoyo a lo largo de este viaje, y que espero se mantengan ahí conmigo en el próximo que emprenda. Desirée, Conchi, Dani, Marina..... gracias.

Por otro lado, la realización de este proyecto no habría sido posible sin el apoyo de mi tutor, Sergio Sanz, ni el de resto de personas con las que he trabajado y he iniciado una nueva etapa; entre todos me habéis y estáis aportando valores y conocimientos que debo agradecer también. Sergio, gracias por la dedicación y la ayuda ofrecida a lo largo de este periodo de tiempo.

Por supuesto, no puedo finalizar sin antes agradecer a mis compañeros de clase todos los momentos únicos e inolvidables que hemos compartido, pues, durante las horas y horas que hemos estado juntos, hemos llorado, reído, enfadado y reconciliado. Gracias a Juanlu, Ana, Cris, David, Marian, Álvaro, Laura, Dani (Talavera), David (Fuenlabrada), Toni, Camarma, y, en definitiva a toda la “Familia

Sonimágica" con la que he pasado grandes momentos, pues todos y cada uno de vosotros me habéis aportado muchísimo y estoy muy agradecida por haberos conocido.

Para terminar, necesito agradecer todo el apoyo, cariño y comprensión recibido por parte de alguien muy especial con quien he tenido la suerte de cruzarme en este camino y que desde entonces es una de las personas más importantes que me rodea. Gracias Javi.

Resumen

En este Proyecto Fin de Carrera se propone un sistema de segmentación del movimiento en secuencias de vídeo y su aplicación a la codificación de vídeo basada en consideraciones perceptuales, uno de los campos de trabajo más interesantes dentro del marco de la codificación de vídeo.

Considerando que la mayoría de secuencias de vídeo cuenta con unas regiones en las que el ojo humano fija su atención (por ejemplo, un objeto que se mueve en un fondo estático), la codificación de estas secuencias se puede llevar a cabo de una manera inteligente asignando más recursos (o bits objetivo) a dichas regiones de interés para que la secuencia decodificada sea subjetivamente más agradable. En función de las principales características del sistema visual humano, que son analizadas en este proyecto, se recurre a un análisis plano a plano del contenido de vídeo para determinar qué regiones son susceptibles de introducir una mayor distorsión de codificación, para que ese ahorro de bits pueda asignarse a las regiones verdaderamente de interés. Estamos hablando, por ejemplo, de la capacidad de segmentar un plano según sus texturas y su cantidad de movimiento.

Concretamente, esta propuesta se centra en la cantidad de movimiento característica de áreas del plano para delimitar las regiones de interés. Una vez se dispone de la segmentación correspondiente, la cuantificación llevada a cabo en el codificador de vídeo será incrementada solamente en aquellas partes del plano cuya calidad pueda ser degradada sin repercusiones subjetivas importantes.

Por último, en este proyecto se incluye un conjunto de pruebas subjetivas que fueron llevadas a cabo para evaluar dos versiones del sistema propuesto de segmentación del movimiento, una básica y una mejorada, implementadas en el

codificador de vídeo H.264/AVC. Un análisis detallado de dichas pruebas así como una serie de propuestas futuras serán expuestos en la parte final de la memoria.

Palabras clave: perceptual, codificación de vídeo, movimiento, segmentación, enmascaramiento, cuantificación.

Abstract

This thesis proposes a motion segmentation system in video sequences and its application in video coding based on perceptual considerations, one of the most interesting fields of work of video coding.

Most of video sequences has some regions where the human eye focuses its attention (i.e, object moving on static background), the codification of these sequences can be carried out in an intelligent way allocating more resources (target bits) in the regions of interest so that the decoded sequence is subjectively more pleasant. Depends on the main characteristics of human visual system that have been analyzed in this project a video content study is made frame to frame to find the areas most susceptible to introduce more distortion coding, so these save bits can be applied to the regions of interest. All of this is based on the capacity of segment a frame from its texture or movement.

Specifically this proposal focuses on amount of movement of frame areas in order to segment the regions of interest. When the segmentation is done, the video coder quantification is increased in areas where the quality can be degraded without important subjective penalties.

Finally, this thesis includes a group of subjective tests were realized to evaluate two different versions of the proposed motion segmentation system: a basic version and other one improved, both of them implemented in video coder H.264/AVC. A detailed analysis of these tests and some future proposals are exposed at the end of this document.

Keywords: perceptual, video coding, motion, segmentation, masking, quantification.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y objetivos

1.1 Motivación del proyecto

En el ámbito de la codificación, la señal digital de vídeo está sometida a una serie de distorsiones, generadas durante los procesos de adquisición del vídeo, compresión, procesado, almacenamiento, transmisión y/o reproducción. En concreto, los sistemas de compresión de vídeo con pérdidas pueden degradar la calidad durante la cuantificación, pues se basan en el principio de eliminar la redundancia espacio-temporal del contenido de vídeo y cuantificar la energía del error de predicción, de modo que la información que primero se pierde es, en general, la alta frecuencia (detalles y bordes). Desde un punto de vista subjetivo, esta información es la menos relevante debido a la "torpeza" del ojo de apreciar la información de detalle desde una cierta distancia al dispositivo de presentación. Sin embargo, se considera que no se están explotando del todo las propiedades del ojo humano para codificar el contenido de vídeo según criterios perceptuales.

Si bien las técnicas de codificación perceptual de audio han sido desarrolladas a partir de extensos estudios psicoacústicos sobre cómo el oído humano y el cerebro detectan e interpretan el sonido; de la misma manera, la codificación de vídeo debería requerir un estudio más exhaustivo acerca del

sistema visual humano, para determinar las características del mismo que influyen en la percepción subjetiva de la calidad.

Existen numerosos estudios realizados con respecto al complejo proceso de percepción visual humana que pretenden determinar en una secuencia de vídeo aquellas regiones que suscitan mayor interés al observador. Por consiguiente, se han elaborado métodos de introducción de distorsión localizada en aquellas regiones catalogadas como de poco interés subjetivo, en base a la respuesta del sistema visual humano. Por lo general, esos métodos tratan de realizar, una asignación inteligente de bits objetivo a cada unidad de codificación atendiendo a criterios de percepción humana.

En la tarea de codificación se dispone de un cierto presupuesto de bits por plano que se pretende emplear en dicho proceso, de modo que el reparto del mismo debe llevarse a cabo con el fin de que la calidad final de la secuencia codificada sea aceptable. Al introducir consideraciones perceptuales en la codificación, dicha asignación de bits debería depender de las limitaciones que el sistema visual humano (HVS, *Human Visual System*) presenta. Los límites del HVS permiten determinar zonas en el plano, denominadas *de atención*, a las que se les asigna mayor número de bits por focalizar en ellas la atención del observador; por otro lado, en aquellas zonas consideradas de menor importancia por el HVS se aplican técnicas de enmascaramiento que consiguen una reducción en la asignación de bits correspondiente y, como consecuencia, una reducción de la calidad visual de esas zonas.

Existen dos vertientes principales encargadas de abordar esta tarea:

- *Regiones de interés*. Son aquellas de las que el ojo humano está más pendiente, de modo que es necesario realizar una segmentación de cada plano donde se diferencien estas regiones de las menos significativas; cada región se codifica de la manera más adecuada. Este método tiene algún inconveniente y es que no es del todo aconsejable cuando las regiones de mayor interés no están bien definidas; por ello, este método quedaría restringido a un conjunto de secuencias predeterminadas como por ejemplo las correspondientes a una videoconferencia o a algún deporte concreto, donde el centro de atención puede detectarse con cierta facilidad.
- *Enmascaramiento visual*. En esta vertiente también se realiza una selección de regiones de interés, pero en este caso, se buscan zonas del

plano donde los errores son menos perceptibles, es decir, el objetivo es encontrar aquellas regiones en las que se puede introducir distorsión sin que el ojo lo aprecie. Para este modo de codificación es elemental el estudio de las características del sistema visual humano, porque determinan las bases para delimitar las zonas en las que la degradación de la imagen es más perceptible, y así, posteriormente, aplicar a las mismas una distorsión apropiada.

Es en la segunda vertiente donde se sitúa el sistema propuesto del actual proyecto, pues a diferencia de la otra vertiente, este método es aplicable independientemente del tipo de secuencia, pues la atención visual del sujeto no rige la segmentación, sino las características de movimiento y/o textura de la escena. Por tanto, como consecuencia del estudio de técnicas actuales en el marco de la codificación perceptual de vídeo, el sistema presentado en este documento pretende introducir su aportación en esta línea de investigación aún en desarrollo.

1.2 Objetivos

El objetivo principal del proyecto consiste en la implementación de un sistema de segmentación del movimiento en secuencias de vídeo y su posterior evaluación integrándolo en un codificador de vídeo H.264/AVC.

Basándose en la cantidad de movimiento presente en la escena, el sistema determina qué regiones son más susceptibles de enmascaramiento, dando lugar a una segmentación de cada plano en áreas con poco y mucho movimiento. Teniendo en consideración la respuesta del sistema visual humano frente al movimiento en una secuencia, sobre un objeto del plano que presente un movimiento elevado se podrá aplicar más distorsión que sobre aquellos que se muevan más despacio. Sin embargo, el foco de interés para el observador no siempre es aquel objeto que presente un movimiento elevado, pues influyen factores adicionales, como la situación espacial en el plano de dichos objetos, o las características ambientales del lugar donde se visualiza el vídeo, entre otros; por tanto, la evaluación del sistema propuesto determinará la necesidad de algún módulo de procesado adicional que complemente la clasificación obtenida.

Por otro lado, la técnica de aplicación localizada de distorsión a la que se recurre consiste en el incremento del parámetro de cuantificación (QP,

Quantization Parameter) del codificador H.264/AVC, siendo más alto cuanto menos sensible al enmascaramiento sea cada región, en función de la combinación de clasificaciones obtenida. En referencias bibliográficas consultadas, se han encontrado técnicas similares de distorsión localizada pero basadas en otras magnitudes perceptuales características de las secuencias.

Este trabajo pretende combinar esta técnica de incremento de QP con el enmascaramiento por movimiento que puede producirse en una secuencia, con el fin de conseguir una mejora de la calidad subjetiva del resultado y una reducción de la tasa de bits utilizada en la codificación, gracias al reparto inteligente de bits del presupuesto en cada región del plano.

Una vez implementado el sistema de enmascaramiento por movimiento es necesario evaluar la viabilidad del mismo, por medio de pruebas subjetivas de codificación, para comprobar la calidad visual del resultado final y la posible reducción de tasa conseguida. El análisis de resultados de estas pruebas es determinante para los trabajos y modificaciones posteriores.

1.3 Estructura de la memoria

A continuación, se describen las partes que conforman la memoria de este proyecto, y se realiza una breve descripción del contenido detallado en cada una de ellas.

En primer lugar, el documento consta de 8 capítulos, de entre los cuales, este primero consiste en una introducción general a toda la información que está contenida en la memoria, así como de los motivos que han dado lugar a la realización de este proyecto y los objetivos que éste pretende alcanzar.

El capítulo contiguo incluye una presentación detallada del estado del arte de las consideraciones perceptuales aplicadas a codificación de vídeo; en éste se presentan trabajos de investigación realizados al respecto, así como ciertas técnicas destacadas que sirven de base en la construcción del proyecto.

A continuación, el tercer capítulo es una descripción del sistema completo propuesto, que expone la idea general del trabajo realizado, sin entrar en detalle en las técnicas o algoritmos utilizados. La tarea de describir en detalle el algoritmo

de enmascaramiento por movimiento es descrita en el capítulo 4. Aquí, además de describir el funcionamiento del mismo, se incluye la configuración de los parámetros pertinentes, así como una primera versión del clasificador de movimiento.

El capítulo 5 es más denso que el resto pues detalla dos etapas de procesado añadidas al sistema cuyo objetivo es mejorar el mapa de vectores característico del movimiento y, consecuentemente, la clasificación final. También consta de un apartado de mejora del clasificador de movimiento.

Para evaluar la viabilidad del sistema desarrollado se realizan dos pruebas subjetivas de codificación basada en consideraciones perceptuales, cuyos resultados se recogen en el capítulo 6.

El capítulo 7 aporta las conclusiones obtenidas del análisis de resultados correspondientes a las pruebas realizadas y, adicionalmente, se incluye un apartado de posibles mejoras a realizar en el sistema propuesto. En definitiva, partiendo de las observaciones realizadas, se exponen algunas ideas que conforman las posibles líneas futuras de trabajo.

Por su parte, el capítulo 8 detalla el presupuesto estimado de la realización de este proyecto, desglosando los costes asociados al material y a los honorarios correspondientes.

Para finalizar, es necesario destacar la existencia de un conjunto de secuencias anexas a esta memoria que actúan como ayuda complementaria para la comprensión de las decisiones tomadas a lo largo del trabajo y que suponen los resultados de las pruebas llevadas a cabo en cada etapa de desarrollo; a ellas se hace referencia a lo largo del texto como apoyo visual de las conclusiones extraídas en cada caso. Todas las secuencias están recogidas en el DVD adjunto, cuya organización en directorios se detalla en el anexo añadido al final de este documento. Adicionalmente, se presenta un glosario que recopila todos los acrónimos utilizados en numerosas ocasiones a lo largo de todo el texto para facilitar la lectura del mismo.

Capítulo 2

Estado del arte

2.1 El Sistema Visual Humano (HVS)

El objetivo de diseño principal en codificación de vídeo o imagen es que el producto final que ofrece el sistema sea aceptable y agradable a la vista para el observador que puede evaluar el resultado. Para la consecución de este fin es necesario tener en cuenta la respuesta del sistema visual humano ante diferentes impulsos luminosos, así como la interpretación de ciertos estímulos.

Además, el factor psicofísico es importante e influyente en la percepción de un vídeo o imagen, determinante en la selección de los objetos de interés por parte del observador. Por ello, se requiere incluir ciertas nociones acerca de la atención visual, que desarrollen el proceso de seguimiento de los ojos (*eye-tracking*) y la motivación del movimiento ocular.

En primer lugar, es necesario conocer la anatomía del ojo y su funcionamiento para extraer conclusiones aplicables en el desarrollo de técnicas de codificación. La referencia [12] es una muestra de dicha necesidad, pues incluye una descripción del proceso de percepción visual para conseguir un modelo de

codificación basado en la fóvea, región de la retina del ojo humano cuyas características se detallan posteriormente.

En el sistema visual humano se distinguen tres componentes principales: los ojos, las áreas visuales del cerebro y las trayectorias visuales. En este apartado del proyecto se aportan conocimientos genéricos de cada uno de los componentes para facilitar la comprensión del uso de ciertas técnicas en el desarrollo de sistemas de codificación.

2.1.1 El ojo humano

El primero de los componentes destacados en el HVS es el ojo. Una sección transversal del mismo se presenta en la figura, en la que son apreciables las diferentes partes de las que se compone, entre ellas, las más destacadas funcionalmente que se comentan a continuación y que describen [18] y [23].

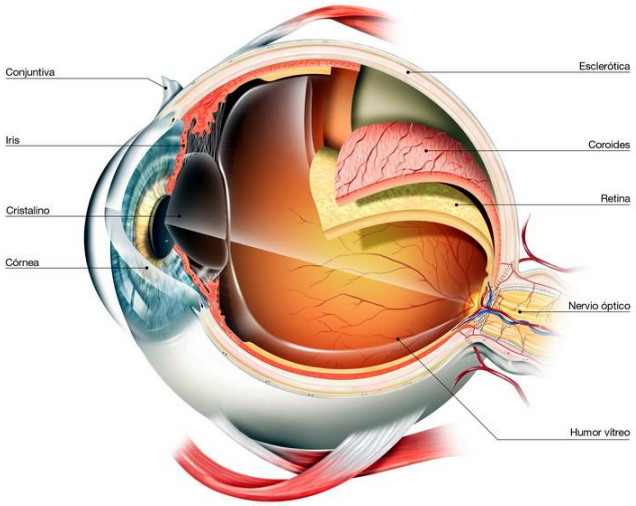


Figura 2.1. Ojo humano.

Los objetos emiten o reflejan radiaciones luminosas de distinta frecuencia e intensidad que entran en el interior del *globo ocular* a través de la *pupila*. El *iris* controla la apertura de la lente y por lo tanto, la cantidad de luz que entra en el ojo. La luz es enfocada e invertida por la *córnea* y el *cristalino*, y es proyectada sobre la ***retina***, la parte fotosensible del ojo.

La retina es una capa muy fina de células neuronales responsables de convertir las señales luminosas en señales neuronales, y después guiarlas al cerebro a través del *nervio óptico*; el lugar donde parte el nervio óptico se denomina *punto ciego*,

cuyo nombre hace referencia a la ausencia de células fotorreceptoras en dicho punto. Existen dos clases de fotorreceptores presentes en la retina:

- ***Bastones***: extraen información de luminancia a bajos niveles de luz, responsables de la *visión escotópica*.
- ***Conos***: son fotorreceptores de color a altos niveles de luz (*visión fotópica*); perciben mejor los detalles y más rápido los cambios en imágenes que los bastones. En cuanto a su tiempo de respuesta al estímulo, los conos son más rápidos que los bastones. Existen tres tipos de conos, sensibles a la luz roja, verde y azul, que tienen respuestas en frecuencia distintas. En la figura se observan las curvas de sensibilidad de cada tipo de cono, la correspondiente al receptor azul es discontinua pues ha sido escalada para poder apreciar la comparación [23].

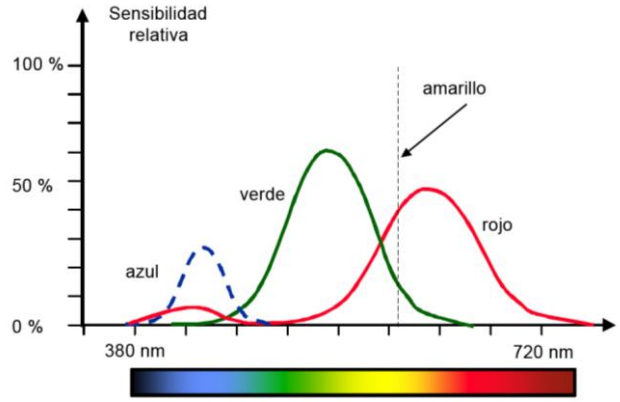


Figura 2.2. Curvas de sensibilidad de los conos.

El sistema visual humano es variante en el espacio debido a que la retina no tiene una densidad uniforme de células fotorreceptoras, tal y como se observa en la Figura 2.3 (correspondiente a la Figura 1.8 de [23]), pues es la ***fóvea*** la zona que contiene la mayor densidad y donde se concentran los conos mayoritariamente; así, cuando el estímulo visual se proyecta en la fóvea, es percibido a la mayor resolución, de aquí que la fóvea sea la responsable de nuestra agudeza visual. A pesar de que la agudeza visual decrece cuando aumenta la distancia al punto de enfoque, un modelo considerado para compresión de vídeo basado en la fóvea puede permitir una representación más inteligente de la escena visual, consiguiendo mejoras en la calidad subjetiva tras la tarea de codificación [12].

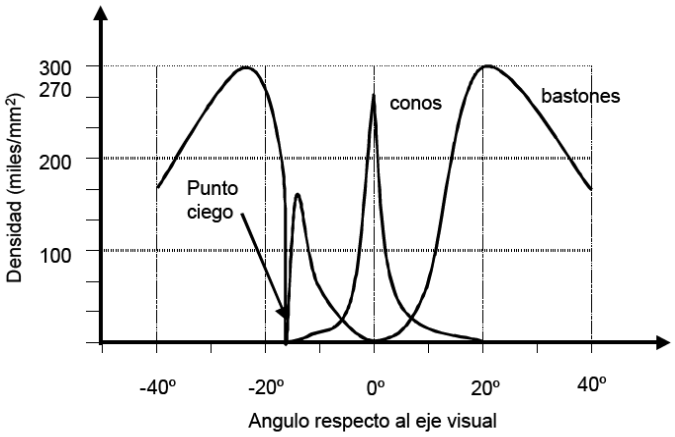


Figura 2.3. Distribución de densidad de conos y bastones en la retina.

2.1.2 Áreas visuales del cerebro

Por otro lado, tras la explicación anatómica del ojo, hay que continuar el camino que recorre el estímulo visual tras abandonar el ojo, por ello nos situamos en los centros visuales del cerebro.

En la percepción visual se encuentra involucrado el proceso fisiológico sufrido por el estímulo en el cerebro. Tras la conversión del estímulo original en señal nerviosa, ésta viaja hasta el *quiasma óptico* a través del nervio óptico, donde se redistribuyen fibras nerviosas; el resultado de esto pasa a los *cuerpos geniculados* que se conectan con las *radiaciones ópticas* (Figura 2.4). Finalmente, las radiaciones ópticas se conectan con los *lóbulos occipitales* derecho e izquierdo de la corteza cerebral, destino final donde se lleva a cabo el proceso psicológico de la percepción. Tal y como se indica en [21] se conocen unas 30 áreas visuales localizadas en los lóbulos occipitales, parietal, temporal y frontal de la corteza cerebral, y que cada área extrae diferente tipo de información correspondiente a la entrada visual.

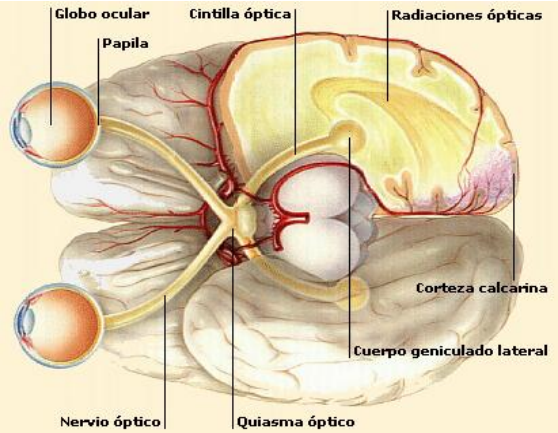


Figura 2.4. Áreas visuales del cerebro implicadas en el proceso de percepción visual.

Para terminar con este componente influyente del HVS, señalar que, aunque el acto perceptivo tenga lugar de forma automática, es realmente complejo y tiene múltiples implicaciones. Según [30] la interpretación de la información recibida varía según el individuo, viéndose alterada por la cultura, la educación, la edad, inteligencia, e incluso estado emocional.

2.1.3 Trayectoria visual

Otra componente destacada del sistema visual humano es la trayectoria visual o camino característico del movimiento que realizan los ojos durante el proceso de percepción o atención visual.

El nivel de atención puede variar mucho y puede estar influenciado por muchos factores como el tamaño, la forma, el color, la localización de un objeto determinado… Normalmente, las figuras humanas y las caras suelen ser el centro de atención; sin embargo, una imagen puede no ser retenida con un simple vistazo ya que la retina es muy variante espacialmente en el procesamiento y muestreo de la información visual. Por lo tanto, los ojos se mueven buscando las partes destacadas de una escena para reconstruir la realidad observada.

El ojo humano hace movimientos oculares rápidos, denominados *sacádicos*, debido a que únicamente la parte central de la retina, la fóvea, tiene una alta concentración de conos. Estos movimientos hacen que el ojo centre su atención en las partes más destacadas de la imagen, captando diferentes partes de una misma escena, de ello se benefician las aplicaciones de vídeo o imagen asignando mayor tasa de bits a las zonas salientes de la imagen consiguiendo mejor calidad visual. Generalmente, los movimientos de los ojos en el proceso de atención visual

realizado se denominan *bottom-up* y *top-down*, siendo el primero realizado de forma involuntaria por los ojos, mientras el otro se realiza intencionadamente; estos dos tipos de movimientos serán explicados de manera detallada en el apartado 2.2.3.3, correspondiente al enmascaramiento causado por el movimiento y la visión temporal.

Gracias a experimentos de seguimiento ocular (*eye-tracking*) desarrollados se consiguen patrones de mirada que identifican las zonas de mayor interés visual, información que será utilizada en la asignación de mayor número de bits a dichas zonas, mientras se degrada la calidad del resto de la imagen. El trabajo [19] recoge las dos técnicas para monitorizar *eye-tracking*, la primera es aquella que mide la posición del ojo relativa a la cabeza, y por otro lado, la que mide la orientación del ojo en el espacio.

Por otro lado, el desarrollo de modelos de atención visual (VA, *Visual Attention*) puede ser beneficioso en aplicaciones de vídeo o imagen para identificar regiones de interés (ROI, *Region of Interest*) que serían tenidas en cuenta en asignación de calidad. Sin embargo, el factor subjetivo está presente en el diseño de estos modelos, pues dependen altamente de los datos subjetivos utilizados como base. Existen modelos de atención visual, como los obtenidos en experimentos de *eye-tracking*, desarrollados para la validación y el diseño de modelos; una muestra de ello sería [13], donde se realiza una comparación de dos modelos de VA: ROI selectiva y patrones de fijación visual (VFP, *Visual Fixation Patterns*). El segundo método recurre al TM3 *eye-tracker*, equipado con dos fuentes de luz infrarroja y una cámara infrarroja para seguir la mirada de ambos ojos, herramienta con la que se concluye que las caras humanas, y sus ojos especialmente, son el foco de atención de los observadores, y además permite determinar la mayor precisión del modelo VFP con respecto a ROI selectiva.

2.2 Consideraciones perceptuales aplicadas a codificación de vídeo

2.2.1 Principios básicos de codificación de vídeo

El escenario en el que normalmente se incluyen las consideraciones perceptuales es la codificación de vídeo y, en particular, el estándar de codificación

H.264/AVC el más reciente y el que mayor eficiencia de compresión consigue. Es necesario conocer el funcionamiento básico de este codificador para tener conciencia del ámbito de codificación al que hay que acondicionar las técnicas de enmascaramiento perceptual a desarrollar. Así, está destinado a funcionamiento en tiempo real, lo que aporta ciertas restricciones a la hora de codificar; también, su aplicación es genérica, en relación con el tipo de secuencias a codificar y se busca que la complejidad esté limitada. Las aplicaciones a las que está destinado este estándar son: *broadcast*, *streaming*, almacenamiento de vídeo y playback (DVD), videoconferencia, vídeo móvil y distribución de estudio; cada una de ellas presenta una serie de requisitos que hay que cumplir como puede ser, por ejemplo, baja latencia en videoconferencia y en vídeo móvil.

En primer lugar, es necesario conocer los conceptos generales de compresión presentes en las diferentes etapas de las que consta el codificador de vídeo H.264, así como las novedades que aporta. Los formatos de compresión de vídeo o CODECs consisten en una pareja COdificador/DECodificador, categorizada como con pérdidas (*lossy*) cuando la imagen reconstruida presenta distorsión y pérdidas de información, o, sin pérdidas (*lossless*), si se obtiene una imagen exactamente igual a la original. En este caso se trata de un codificador con pérdidas cuyo procesado va destinado a la eliminación de redundancias en los datos, lugar donde tienen cabida las consideraciones perceptuales. Un esquema general del CODEC H.264 sería el mostrado por las Figuras 2.5 y 2.6 (extraídas de [20]), donde se aprecian los diferentes bloques que lo componen y que se detallan posteriormente.

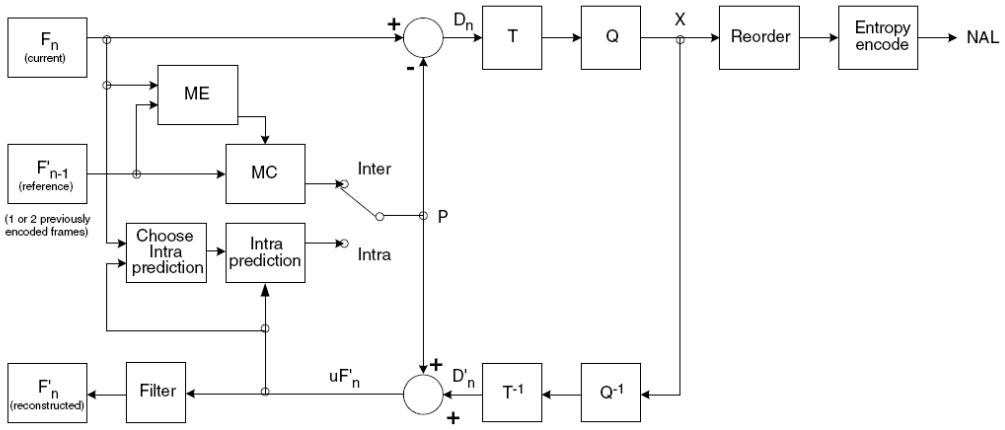


Figura 2.5. Codificador H.264/AVC.

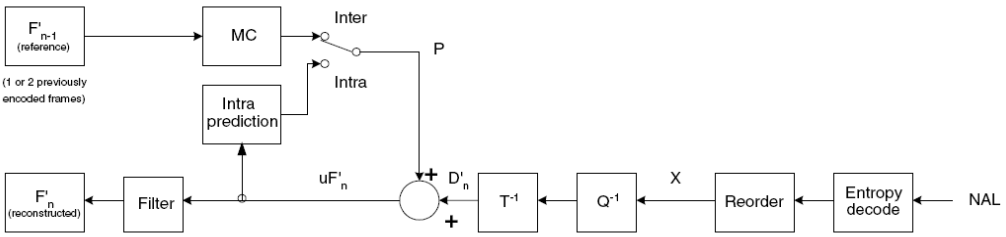


Figura 2.6. Decodificador H.264/AVC.

Como todo formato de compresión, el procesado realizado para eliminar la redundancia se puede dividir en tres partes o modelos: temporal, espacial y estadístico, según qué tipo de redundancia se trate. Haciendo uso de estos tres modelos, se detallan a continuación los bloques que componen el codificador implicado en el proyecto.

- **Modelado temporal**

Esta etapa se encarga de la caracterización del movimiento de la secuencia de vídeo mediante el cálculo de un mapa de vectores de movimiento y una imagen residual. Las fases del modelado temporal son las siguientes:

(1). *Estimación del movimiento*: tomando como referencia un bloque de una o varias imágenes pasadas o futuras ya codificadas, elegidas de manera que cumplan un cierto criterio.

(2). *Compensación del movimiento:* diferencia entre el bloque original y el correspondiente en la imagen de referencia, generando un bloque residual. Una novedad introducida en H.264/AVC con respecto a los anteriores estándares de codificación es la posibilidad de realizar la compensación con múltiples imágenes de referencia.

(3). *Cálculo del residuo:* diferencia entre la imagen original y la imagen predicha con los vectores. Este residuo es almacenado junto a la imagen predicha como imagen de referencia.

El estándar H.264/AVC considera el macrobloque (MB) de 16x16 la unidad básica para la predicción de movimiento compensado, aunque a veces se recurre a tamaños menores (8x8 o 4x4) para conseguir una predicción más acertada. Hay que destacar una técnica muy utilizada, la compensación sub-píxel, para poder detectar movimientos inferiores a un píxel. Se recurre a

técnicas de interpolación sobre el macrobloque en estudio y la imagen de referencia; esto supone un coste computacional elevado y un aumento en los bits correspondientes con los vectores, a pesar de la reducción de energía del residuo que se consigue, siendo necesaria una solución de compromiso, pero se consigue una precisión de vectores de hasta un cuarto de muestra.

Entonces, existen tres tipos de predicciones para estos MB:

- *I*: se predicen mediante predicción *Intra*, es decir, sin compensación de movimiento y a partir de muestras previamente decodificadas del cuadro actual. Se predice el MB completo o cada sub-bloque de muestras del MB.
- *P*: utiliza predicción *Inter*, a partir de una o más imágenes de referencia almacenadas en dos listas denominadas *list0*, que almacena imágenes pasadas, y *list1*, que almacena imágenes futuras. En este caso, el MB puede ser dividido en particiones, es decir, sub-bloques de 16x16, 16x8, 8x16 o 8x8; a su vez las particiones se pueden dividir en sub-particiones de 8x4, 4x8 o 4x4.
- *B*: utiliza predicción *Inter*, tomando dos referencias, una de *list0* y otra de *list1*.

Además, con respecto al formato del plano del codificador H.264/AVC, hay que destacar la existencia de un nivel superior al MB, se trata de las *Tiras* o *Slices*, conjuntos de MBs; su tamaño es flexible. Pueden ser de tipo I, si solo contienen MB de este tipo; P, si lo forman MBs I y P, o tipo B que pueden estar formadas por cualquier tipo de MB. Además, existen las tiras tipo SI y SP para conmutación entre flujos de vídeo diferentes. Para conseguir mayor robustez ante errores se utilizan los grupos de tiras (FMO, *Flexible Macroblock Ordering*). El envío de las tiras de un plano puede ser en orden arbitrario (ASO, *Arbitrary Slice Ordering*), mejorando el retardo extremo a extremo.

Por último, por encima de las tiras está el nivel de cuadro (*frame*) para vídeo progresivo, y el nivel de campo (*field*) para vídeo entrelazado. Y, a nivel global, la secuencia de vídeo se organiza en grupos de planos (GOP, *Group Of Pictures*), de diferentes tipos (I, P, B) que siguen un patrón concreto. El plano I encabeza el conjunto sin ninguna referencia anterior y luego le siguen los planos P y B, de modo que la imagen I sirve como referencia a P y B, mientras la imagen P se utiliza para predecir otras P y B, mientras las B no suelen ser utilizadas como referencia, a excepción del estándar H.264/AVC que recurre a

planos tipo B como planos de referencia para conseguir una mayor compresión (véase la Figura 2.7).

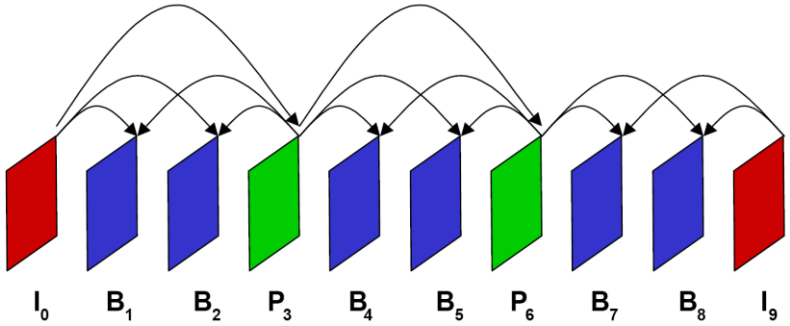


Figura 2.7. Referencias en GOP IP2B.

En caso de que el patrón contenga planos I y P (denominando IP a este tipo de patrón), el orden de codificación de los planos es el mismo que el de presentación. Por su parte, para codificar un plano P cuando el patrón incorpora planos B (bidireccionales) (IPXB) el orden que siguen es el siguiente:

- Orden de presentación: $I_0\ B_1\ B_2\ B_3\ P_4\ B_5\ B_6\ B_7\ P_8$...
- Orden de codificación: $I_0\ P_4\ B_1\ B_2\ B_3\ P_8\ B_5\ B_6\ B_7$...

- **Modelado espacial**

Esta etapa parte del residuo calculado en el modelado temporal y busca explotar las redundancias espaciales del mismo.

Para ello, utiliza, en primer lugar, codificación predictiva (DPCM, *Differential Pulse Code Modulation*), calculando predicciones de muestras a partir de muestras vecinas ya codificadas, explotando correlaciones locales; en concreto, los vectores de movimiento de bloques vecinos son similares, por ello se predice el movimiento en base a vectores previamente calculados. La información que se transmite en este caso es el vector de movimiento diferencia (MVD, *Motion Vector Difference*), que es la diferencia entre el vector real y el predicho. Además de en los vectores de movimiento, la codificación predictiva está presente en el escalón de cuantificación (QP), pues varía poco de un plano al siguiente.

Por otro lado, en el modelado espacial interviene la codificación transformada, para decorrelar las muestras, con el fin de conseguir una tasa de compresión mayor. Las dos transformadas más destacadas son: *Transformada DCT o Transformada Discreta del Coseno* y la *Transformada Wavelet*, pero se recurre a la DCT, que se detalla a continuación.

La *Transformada DCT* opera a nivel de bloque (NxN) y es reversible (existe IDCT), además en el caso de H.264/AVC la transformación inversa es exacta. Proporciona una matriz de coeficientes que son los pesos de una base de NxN funciones. De todos los coeficientes obtenidos sólo los más próximos al coeficiente (0,0) disponen de una energía significativa, teniendo el resto, en general, una energía despreciable. Los coeficientes se ordenan realizando un escaneo en zigzag del MB, que varía según se trate de vídeo progresivo o entrelazado, según indican las Figuras 2.8 y 2.9.

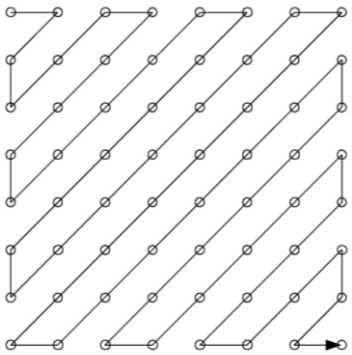


Figura 2.8. Escaneo en zigzag (de cada MB del plano).

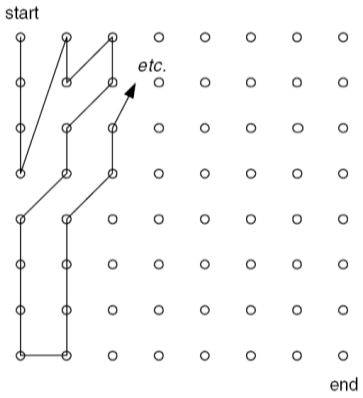


Figura 2.9. Escaneo en zigzag (de cada MB de cada campo).

Los patrones base para DCTs de 4x4 y 8x8 están representados en la Figura 2.10. La transformación en el estándar es jerárquica, puesto que se pueden utilizar diferentes tamaños de bloque en función del contenido.

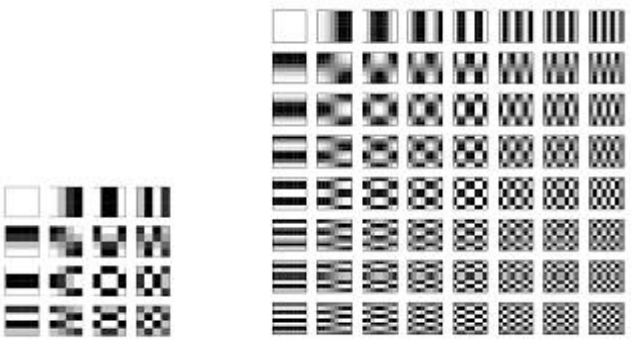


Figura 2.10. Patrones base para DCTs de 4x4 (izquierda) y 8x8 (derecha).

Como se puede observar en la Figura 2.5, correspondiente al codificador H.264/AVC, tras la etapa de transformación, T, aparece un bloque Q o de *cuantificación*. La cuantificación es el proceso en el que se introducen las pérdidas de información en el sistema; consiste en la conversión de una sucesión de muestras de amplitud continua en una sucesión de valores discretos. La cuantificación más sencilla es la *cuantificación uniforme*, cuya función de transferencia es la siguiente:

$$FQ = round\left(\frac{X}{QP}\right) \qquad (1)$$

$$Y = FQ \cdot QP \qquad (2)$$

FQ es el valor cuantificado, a partir del valor de entrada X; por otro lado, Y se corresponde con el valor reconstruido. El parámetro QP es la distancia entre dos valores cuantificados consecutivos; por lo tanto, es el encargado de determinar el grado de compresión alcanzado.

Otro tipo de cuantificación es la *vectorial*, cuyo nombre procede del tratamiento que se realiza de los datos en forma de vector. En este caso, no se cuantifican las muestras individualmente, sino en bloques de muestras o vectores.

- **Modelado estadístico**

Después de los modelos temporal y espacial, se obtiene un conjunto de datos que presentan patrones estadísticos y que constituyen todo el flujo de

información que se va a transmitir o almacenar. El objetivo de este bloque es reducir las redundancias estadísticas presentes en esos datos para conseguir una mayor compresión del resultado final, de ello se encarga la *codificación entrópica*. Existen dos tipos de códigos utilizados en el estándar H.264/AVC, denominados *CAVLC* (*Context-Based Adaptive Variable Length Coding*) y *CABAC* (*Context-based Adaptive Binary Arithmetic Coding*) y que se explican a continuación.

1. ***CAVLC,*** codificación de longitud variable, en general, asignación de palabras clave cortas a símbolos de entrada muy frecuentes, y más largas para aquellos menos frecuentes.

   Codificación residual de los coeficientes transformados de un bloque 4x4 (y 2x2) ordenados en zig-zag. Estos coeficientes contienen gran número de ceros; también, los coeficientes de alta frecuencia suelen ser secuencias de ±1 que CAVLC indica de manera compacta, y el número de coeficientes distintos de 0 se codifican utilizando una *look-up table*. Para codificar la magnitud de un coeficiente distinto de cero eligiendo un VLC en la tabla se tiene en cuenta que ésta disminuye a medida que nos alejamos de la componente DC.

2. ***CABAC***, codificación aritmética binaria.

   En esta codificación, en primer lugar, se realiza una binarización de los datos, después se selecciona un modelo de contexto, es decir, un modelo de probabilidad para uno o más *bins* de los símbolos binarizados que se elige dependiendo de unas estadísticas de los símbolos codificados previamente. Una vez que se dispone del modelo de probabilidad se realiza la codificación aritmética, y, para terminar, se actualiza la probabilidad del modelo.

Para terminar con el estándar H.264/AVC es importante añadir una serie de especificaciones que incluye. Para determinar las diferentes funcionalidades recurre a unos *perfiles* establecidos, que proporcionan diferentes grados de calidad y/o versatilidad a la aplicación correspondiente. Inicialmente definieron tres: *Baseline*, *Main* y *Extended*; actualmente, existen 11 perfiles diferentes orientados a diferentes tipos de aplicaciones.

2.2.2 Fundamentos de control tasa en codificación de vídeo

La variabilidad inherente de la información de vídeo implica que el codificador de fuente en general produzca una tasa instantánea de bits variable, la cual debe ser controlada con el fin de ajustarse a la tasa nominal de la red de transmisión, que puede ser un canal de tasa constante (CBR, *Constant Bit Rate*) o variable (VBR, *Variable Bit Rate*). En la codificación CBR, típicamente utilizada en aplicaciones de bajo retardo como videoconferencia, se requiere una adaptación instantánea de la tasa nominal con el fin de asegurar el bajo retardo. Sin embargo, en la codificación VBR, normalmente pensada para aplicaciones de *broadcasting* o *streaming*, una adaptación más a largo plazo de la tasa nominal es factible, a costa de un mayor retardo de transmisión, con el objeto de mejorar la calidad visual.

La tasa de salida del codificador es controlada por un módulo de control de tasa que todo codificador debe incorporar para poder transmitir a través del canal *bitstreams* bajo ciertas restricciones de retardo y calidad. Por lo general, entre el codificador de fuente y la red de transmisión se emplaza un buffer virtual, que es el encargado de simular el comportamiento del buffer del decodificador y amortiguar las diferencias entre la tasa nominal de la red y la tasa instantánea debida a la variabilidad espacio-temporal del contenido de vídeo. El algoritmo de control de tasa debe mantener el buffer virtual en niveles seguros evitando así que se desborde (peligro de *overflow*) o que se quede vacío (peligro de *underflow*) infrautilizando la capacidad del canal. Estos dos efectos son perniciosos para la correcta decodificación del *bitstream*. Una situación de *overflow* provoca que el plano codificado tenga que ser descartado, y una situación de *underflow* obliga al decodificador a que se mantenga en espera hasta que le lleguen nuevos datos a su correspondiente buffer.

Con el fin de mantener la tasa binaria de salida dentro de los límites impuestos por el buffer virtual sin que la calidad visual se degrade notoriamente, el algoritmo de control de tasa asigna, a partir del estado actual del buffer y de la complejidad de la información de vídeo, la cantidad de bits objetivo más apropiada y el correspondiente valor de QP a cada unidad de codificación, que en el caso del estándar H.264/AVC se trata del macrobloque. Para ello, se emplean modelos analíticos de Tasa-Cuantificación (R-Q, *Rate-Quantization*) que determinan la QP más adecuada a partir de una cantidad de bits objetivo y de la complejidad de la unidad de codificación. Estas funciones R-Q normalmente se derivan del modelado

estadístico de la distribución de los coeficientes de la DCT en función de la cuantificación aplicada. Se puede acudir a la referencia [36] para profundizar con más detalle en todos los aspectos relacionados con el control de tasa para codificación de vídeo.

Resulta evidente que el algoritmo de control de tasa y el análisis de la secuencia según criterios de percepción humana pueden combinarse para implementar un algoritmo de asignación de bits objetivo basado en consideraciones perceptuales que aumente el valor de QP a aquellas regiones del plano menos sensibles a errores de codificación.

2.2.3 Características del HVS

Todo el análisis anatómico del ojo humano y el estudio de técnicas de trayectoria visual, ponen de manifiesto la importancia del observador en la determinación final de la calidad de una imagen, de aquí, la importancia y conveniencia de incorporar consideraciones acerca del HVS en aplicaciones de procesado de imagen y vídeo. Gran parte de los métodos prácticos en codificación perceptual de imagen y vídeo están basados en el concepto de JND (*Just-Noticeable Distortion*) detallado a continuación, pero existen otros métodos basados en la importancia de las diferentes regiones del plano, que atienden a las siguientes propiedades básicas del HVS incluidas en [4] y también en [18].

- **Relación de Contraste**

  La respuesta del ojo a cambios en la intensidad de iluminación es no lineal, y debido a propiedades psicovisuales del HVS, no se pueden percibir variaciones finas de señal. [9] señala que la primera teoría con respecto a la relación entre la sensibilidad del ojo y la intensidad de un estímulo fue propuesta por *Ernst Heinrich Weber*. Considerando un área de luz de intensidad $I+\Delta I$ rodeada por un fondo de intensidad I, se cumple la relación:

$$\Delta I = K \cdot I \qquad (3)$$

  donde ΔI es la distorsión mínima de I que los ojos pueden percibir; se le suele nombrar JND que indica que la relación $\Delta I/I$ es constante, con un valor próximo a 0.02 según [18], pero este resultado no se mantiene a muy bajas o muy altas intensidades, donde la JND aumenta notablemente. Por lo tanto, la expresión general $(I+\Delta I)/I$ representa la relación de contraste, constante en un rango de

intensidades medio. Weber observó que para cada modalidad sensorial (de color, de tamaño, de olor, de sonido...), no sólo para la intensidad luminosa, hay una constante ("K"), que cumple la relación establecida.

Más tarde, según indica [9], Gustav Theodor Fechner extendió la ley de Weber usando un algoritmo para describir la relación de ΔS y S, o distorsión mínima del estímulo y estímulo respectivamente. La ley de Fechner es la relación logarítmica entre el estímulo de intensidad y la sensación de intensidad asociada:

$$I = K \cdot ln(S/S_{min}) \qquad (4)$$

donde I es la intensidad perceptual de S y S_{min} el nivel del estímulo por debajo del cual no se percibe sensación.

Observando la relación de Weber-Fechner se puede concluir que más distorsión puede ser oculta en altos niveles de estímulo que en bajos, lo que permite comprimir más sin que el observador lo perciba. Los modelos JND desarrollados generalmente explotan la visibilidad de la mínima distorsión perceptible y asumen que la agudeza visual es consistente en toda la imagen. Sin embargo, la agudeza visual es variante espacialmente y no todas las partes de la imagen serán proyectadas en la fóvea de la misma manera. El umbral de visibilidad de una región de la imagen varía acorde con su posición proyectada en la retina, de aquí la aparición de modelos FJND (*Foveated Just-Noticeable Distortion*) como [12] que tienen en cuenta las propiedades de la fóvea para conseguir una mejora de la calidad subjetiva de vídeo para una misma tasa de bit.

Estudios psicovisuales han demostrado que la sensibilidad visual del HVS puede ser medida mediante una función de sensibilidad al contraste espacio-temporal denominada CSF (*Contrast Sensitivity Function*). Esta función muestra el contraste requerido para detectar una rejilla parpadeando a diferentes frecuencias espaciales y temporales. Además, la sensibilidad del HVS al contraste depende de la luminancia del fondo como podemos observar en la siguiente figura, perteneciente a [18].

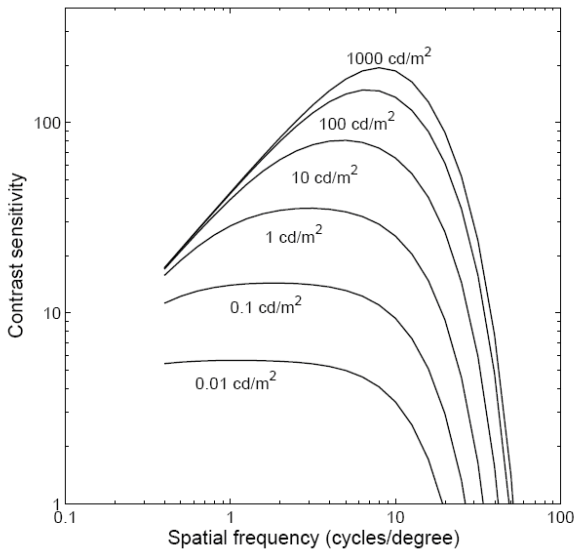


Figura 2.11. Función de Sensibilidad al Contraste para diferentes niveles de luminancia.

- **Frecuencia Espacial**

Otra de las características destacadas del sistema visual humano es la sensibilidad a la frecuencia espacial. En términos de respuesta en frecuencia el HVS actúa como un filtro paso banda; a medida que aumenta la frecuencia la sensibilidad del ojo disminuye, de modo que se puede introducir más ruido en áreas con altas frecuencias espaciales.

El ojo puede percibir un cierto grado de color y detalle cómo máximo, y cualquier detalle que no pueda ser percibido es promediado. Esta propiedad del ojo se denomina integración espacial, y puede ser explotada para eliminar o reducir altas frecuencias sin afectar a la calidad percibida. Por lo tanto, el HVS permite explotar, además de las redundancias temporales y estadísticas, la redundancia espacial.

Existe una correlación significativa entre MBs vecinos de cada plano de una secuencia de vídeo, por ello, explotando las redundancias se puede reducir considerablemente la cantidad de información a codificar. Las técnicas destinadas a ello en codificación suelen ser las transformadas, como la DCT. Las propiedades básicas de las transformadas son la compactación de la energía y la decorrelación, muy adecuadas en compresión de vídeo; así, cuando se aplica una transformada sobre un plano, la energía de cada MB se compacta a unos cuantos coeficientes transformados y la correlación entre dichos coeficientes es

además reducida sustancialmente, lo que implica que una cantidad significativa de información puede ser recuperada utilizando sólo unos pocos coeficientes. La DCT es muy utilizada en codificación de imagen y vídeo porque dispone de una propiedad muy fuerte de compactación de la energía, además de que las funciones base están predefinidas y no es necesario transmitirlas. Los coeficientes transformados en el dominio de la frecuencia son clasificados en baja, media y alta frecuencia, considerado la componente continua al coeficiente (0,0) y a medida que nos desplazamos en vertical y horizontal por la matriz de coeficientes resultante, éstos representan frecuencias espaciales cada vez mayores.

Un ejemplo de codificación perceptual que aprovecha esta propiedad del HVS es [8]. Propone una ponderación de frecuencia a nivel de MB, teniendo en cuenta la sensibilidad a la frecuencia del HVS, se aplica un posible cambio del peso de cada coeficiente de frecuencia sin sacrificar la calidad visual. Las altas frecuencias sufren una ponderación más burda que las bajas, las cuales se corresponden con detalles de la imagen. En caso de que los detalles o bajas frecuencias, de alta sensibilidad subjetiva, se vieran afectadas, la calidad se vería gravemente reducida.

Del mismo modo, [6] recurre a la sensibilidad a la frecuencia espacial del HVS y, como consecuencia, a los coeficientes de la DCT, para obtener una medida de tolerancia perceptual. En concreto, realiza el sumatorio de las frecuencias más bajas de la matriz de coeficientes (coeficientes 1, 2, 8, 9, 10, 16 y 17) como parámetro de medida de bordes prominentes. Así, altos valores de estos componentes indican que el bloque puede que tenga bordes que deben ser conservados.

- **Enmascaramiento**

La visibilidad de la distorsión es un factor muy importante en cualquier aplicación de codificación de imagen o vídeo, por ello éstas aprovechan características básicas del HVS para dejar libres de distorsión a aquellas regiones de un plano/imagen con gran importancia subjetiva.

Cuando la imagen tiene un alto contenido de actividad existe una pérdida de la sensibilidad a errores en las regiones donde se concentra dicha actividad; esto es el efecto del enmascaramiento. Así, el HVS tolera el ruido en texturas del plano aleatorias (según [1]), donde no existen unos patrones prominentes que

destaquen en la imagen y sean del interés del observador. Del mismo modo, si la actividad a la que nos referimos es el movimiento rápido de algún objeto en el plano, aquella región que concentra el movimiento es susceptible al enmascaramiento, teniendo en cuenta la textura que presenta y el grado de movimiento; pues [1] señala que objetos con una textura homogénea que presenta un movimiento regular en el fondo son más sensibles a distorsiones perceptuales, mientras áreas aleatorias en movimiento permiten distorsiones más altas en la codificación. Ambas vertientes del enmascaramiento, tano la relacionada con el movimiento como la de texturas, serán desarrolladas con más detalle en los próximos apartados del actual capítulo, en los que se incluirán ejemplos de técnicas que recurren a dicha propiedad del HVS.

Por otro lado, se han elaborado funciones de visibilidad que miden la sensibilidad del ojo a distorsiones con el fin de cuantificar la cantidad de enmascaramiento y obtener así algún dato numérico en función del cual se actúe en la codificación. Estas funciones de enmascaramiento han sido consideradas en el diseño de cuantificadores para alcanzar baja entropía, según señala [18].

Por último, otro tipo de enmascaramiento puede producirse aprovechando la relación de contraste del HVS pues se puede esconder más ruido en áreas oscuras o muy brillantes de un plano. En sistemas de codificación basados en el dominio de color YUV, los coeficientes de luminancia ocupan la mayor parte de la información visual, por lo tanto, si se consigue reducir la energía de dichos coeficientes se podrá obtener una reducción de la tasa de salida ([9]).

- **Bandas de Mach**

Las bandas de Mach, como se puede observar, consisten en un grupo de barras grises donde la intensidad de cada barra cambia suavemente y de manera constante. Cada banda está separada por una estrecha banda central coloreada con un gradiente de claro a oscuro (Figura 2.12, izquierda). La transición de una a otra es enfatizada por la sensibilidad del HVS a la intensidad, viendo únicamente dos estrechas bandas de diferente luminosidad, tratándose así de una ilusión óptica. El efecto es independiente de la orientación de la frontera.

Esta ilusión óptica se debe al filtrado espacial de alto impulso que realiza el sistema visual humano en la luminancia de la imagen capturada por la retina.

Figura 2.12. Bandas de Mach.

Estas bandas son muy utilizadas para explicar los principales obstáculos que se encuentran en codificación como son los bordes pronunciados en imágenes especialmente en presencia de movimiento, que dan lugar a errores de predicción en codificación predictiva o grandes coeficientes de transformación en codificación transformada.

Como consecuencia del conocimiento de esta propiedad del sistema visual humano en codificación perceptual, se propone introducir distorsión en detalles alrededor de bordes prominentes para impedir la percepción de otras variaciones de contraste más bajas. Sin embargo, el HVS es sensible al desalineamiento de bordes de objetos rígidos, así como a bordes y detalles de zonas con movimiento suave; aunque, percibe menos la distorsión en zonas con alto movimiento y gradiente alto; adicionalmente, en el último apartado relativo al HVS se detallan ciertos fenómenos influyentes en la percepción de movimiento que sufre el HVS.

2.2.4 Tipos de enmascaramiento visual

2.2.4.1 Enmascaramiento por texturas

El enmascaramiento es una de las principales herramientas del sistema visual humano aprovechada en tareas de codificación perceptual de vídeo e imagen. En concreto, este apartado se centra en el enmascaramiento basado en texturas, atendiendo a la sensibilidad del HVS a percibir distorsión en ciertas regiones de la imagen en función de su estructura. La sensibilidad al error es más baja en áreas de la imagen con alto contenido estructurado, lo que permite introducir distorsión en dichas zonas.

Los indicadores del nivel de textura presente en un MB de un plano son los coeficientes de la DCT correspondientes, debido a que éstos tienen una interpretación frecuencial. Así, el resultado de aplicar la DCT sobre un bloque de la imagen es una matriz de coeficientes transformados en el que el elemento (0,0) se corresponde con la componente DC y al desplazarnos en vertical y horizontal por la matriz los valores representan frecuencias espaciales cada vez mayores.

Una muestra de la utilización de los coeficientes DCT como medida de textura serían [4] y [10]. [4] realiza una clasificación de los MB en 6 categorías, teniendo en cuenta también el enmascaramiento por la intensidad de contraste:

1. *Textured*: MB sin patrones estructurados y rápidas fluctuaciones de intensidad. Presentan un espectro plano en frecuencia, en ausencia de frecuencia dominante. Pueden ser codificadas con gran cantidad de ruido.
2. *Dark contrast*: zonas con intensidad mucho más baja que la de las áreas de alrededor.
3. *Smooth*: áreas con intensidad relativamente constante.
4. *Edge*: indica que hay un borde en ese MB.
5. *Detailed*: se tratan de áreas con detalles finos que presentan coeficientes dominantes en la DCT.
6. *Normal*: aquellos que no se determinan en ninguna de las categorías anteriores.

Para determinar la categoría a la que pertenece un MB primero se empieza con bloques de 4x4 dentro del mismo MB y luego con el MB de 16x16. A nivel de sub-bloque (4x4) se calculan las actividades de textura y brillo, mediante la varianza de los coeficientes AC y la media de luminancia del bloque, respectivamente. A nivel de MB se clasifica según la lista anterior; se determina de un tipo o de otro en función de unos umbrales establecidos acerca de las magnitudes medidas en cada sub-bloque.

Por otro lado, [10] incluye un modelo perceptual basado en propiedades de enmascaramiento en texturas y luminancia, pero destinado a un codificador JPEG. En él se recurre también a la energía AC de los coeficientes DCT de un bloque para medir la actividad de textura. Sin embargo, el algoritmo de clasificación, en este caso, consiste en la comprobación de unas condiciones entre la suma absoluta de valores de unos coeficientes concretos y unos umbrales, descrito en detalle en [10]. Los coeficientes se distinguen como indica la figura:

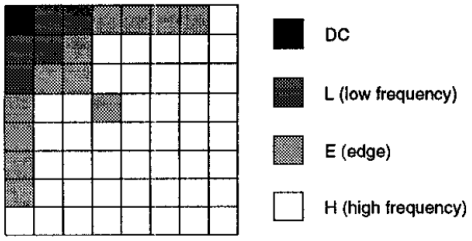


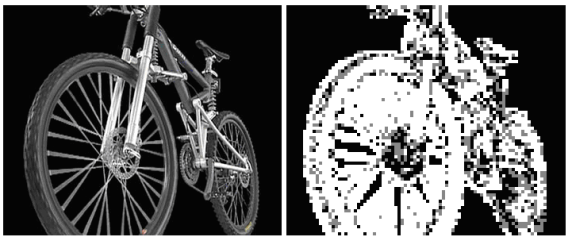
Figura 2.13. Clasificación de bloques.

Otros métodos de análisis de texturas utilizados son descriptores de color, detección de bordes, o filtros de Gabor para extracción de características, en definitiva, herramientas típicas de procesado de imágenes utilizadas para extraer información relevante del plano que poder utilizar posteriormente en enmascaramiento, con el objetivo de reducir la tasa de bits. Mediante este bloque de análisis basado en procesado característico de imágenes se consiguen identificar, entre otras, las zonas homogéneas del plano y éstas a su vez ser etiquetadas como texturas, al igual que aquellas zonas que presentan un patrón estructurado con direcciones prominentes.

En relación con el uso de técnicas de procesado del plano, la referencia [1] hace uso de un operador de detección de bordes aplicable a cada plano de la secuencia a codificar para, posteriormente, evaluar la media de los pesos de los bordes y la distribución de densidad de los píxeles de borde en cada MB; finalmente, procede a la extracción de características locales. Los dos tipos de detectores de bordes estudiados en su caso son Canny y Sobel. El objetivo de este método es el análisis de la sensibilidad visual a la distorsión para un modelo de texturas que discrimina las regiones con texturas aleatorias durante el proceso de asignación de bits. Considera textura aleatoria aquella que contiene pequeños bordes con diferentes orientaciones, mientras una región estructurada se compone de bordes más consistentes y de mayor tamaño. Aunque las texturas aleatorias conllevan más entropía que las estructuradas, el HVS es menos sensible a distorsiones en zonas aleatorias porque éstas encubren los ruidos de codificación. En cuanto a los detectores de bordes utilizados en ese artículo, Canny proporciona una tasa de error muy baja en la detección de bordes pero es más costoso computacionalmente, además, de manera complementaria se recurre al detector Sobel para poder diferenciar las regiones aleatorias y las texturas estructuradas. Para finalizar con dicha referencia, señalar que este modelo de texturas va acompañado de un modelo de atención al movimiento; ambos modelos proporcionan un índice de aleatoriedad de texturas y de atención de movimiento, respectivamente, utilizados en la asignación de QP en el proceso de codificación,

consiguiendo de esta manera que la asignación de bits, y por tanto, la calidad del plano, sea dependiente de las características locales subjetivas extraídas en cada plano.

En otras ocasiones, las texturas se dejan en segundo plano cuando existen bordes prominentes en la imagen, como consecuencia del enmascaramiento producido por los mismos [2]. El HVS es menos sensible a errores a lo largo de un borde prominente, que impedirá la percepción de otras variaciones de contraste más bajas en ese mismo bloque; por lo tanto, mayor tasa de bits se asigna a los bordes patrón, y menos a las texturas caóticas. En el artículo [2], se actualizan los multiplicadores de Lagrange acorde con el patrón del MB, que se clasifica como "*Edge*", "*Texture*" o "*Background*" basándose en los operadores de Sobel característicos del MB, que detecta cambios horizontales y verticales, generando un mapa de importancia del plano como el que muestra la figura.



(a) Imagen Original (b) Mapa de Importancia

Figura 2.14. Ejemplo de segmentación.

2.2.4.2 Enmascaramiento temporal

La sensibilidad del sistema visual humano a cambios en el tiempo y la percepción de objetos en movimiento van unidos a menudo porque el estímulo para el movimiento produce variaciones temporales en la intensidad de luz que cae en la retina. Esto significa que la percepción del movimiento está guiada por los mismos mecanismos que detectan cambios de intensidad a lo largo del tiempo.

La respuesta del sistema visual humano al movimiento se caracteriza por dos hechos: la *persistencia de visión* y el *fenómeno Phi*. Ambos son explotados en televisión y cine para obtener percepción de movimiento a partir de una sucesión rápida de imágenes o cuadros, que se corresponden con la información de una escena particular discretizada en el tiempo. Otro fenómeno que puede producir enmascaramiento temporal es el cambio de escena producido en una secuencia, de

modo que los planos próximos al cambio de escena, tanto anteriores como posteriores, son susceptibles de ser descartados o degradados cualitativamente.

- **Persistencia visual**

Describe la tasa de muestreo temporal del HVS. Este fenómeno consiste en que la percepción de una imagen se mantiene durante unas fracciones de segundo después de que haya desaparecido la excitación; es decir, que la respuesta del ojo a un impulso lumínico no desaparece inmediatamente después del mismo.

Si la frecuencia de excitación es lenta, el sistema visual es capaz de discernir entre distintas excitaciones. Sin embargo, si aumentamos la frecuencia llegará un momento en que se perciba una sensación de iluminación uniforme a lo largo del tiempo; el nivel de iluminación subjetivo que se percibe coincide con el valor medio de la señal utilizada para la excitación, y a este hecho se le conoce como la *ley de Talbot-Plateau*.

En el rango de frecuencias en el que se distingue correctamente la intermitencia del estímulo y el correspondiente a la percepción de continuidad, existe un intervalo en el que se aprecia un parpadeo de la imagen. Con un estímulo intermitente de entre 50-60 Hz se percibe un estímulo uniforme, dependiendo de las condiciones de iluminación y contraste y los observadores; esta es conocida como la *Frecuencia de Fusión Crítica* (*CFF*) o frecuencia *Flicker*. Este valor de frecuencia crítica es común en condiciones de fuerte iluminación diurna, pero puede disminuir hasta los 4 Hz en caso de iluminación nocturna y visión fotópica.

- **Fenómeno Phi**

Por otro lado, el fenómeno Phi describe un umbral por encima del cual el HVS detecta movimiento. Es el responsable de la interpolación de movimientos de los que sólo se dispone información fraccionada, dando la sensación de continuidad. Un ejemplo característico de esta respuesta del sistema visual consiste en dos focos de luz que emiten luz alternativamente con un retardo de frecuencia de unos 62 Hz ([19]), dando la sensación de movimiento aparente; este valor de frecuencia se corresponde con una tasa en torno a 16 fps, de modo que la velocidad crítica se puede considerar de unas 18 imágenes por segundo.

- **Cambio de escena**

Este tipo de enmascaramiento temporal se basa en el cambio brusco de contenidos que sufre el observador al producirse un cambio de escena, donde el plano anterior al cambio y el primero tras el corte tienen características totalmente diferentes.

La referencia [28] demuestra la existencia de un efecto de enmascaramiento temporal asimétrico causado por el cambio de escena. Mediante un experimento de calidad subjetiva se evalúa el impacto de introducir una distorsión temporal, consistente en una ráfaga de planos repetidos, utilizando variedad de contenidos de vídeo, así como diferentes localizaciones y duraciones del deterioro con respecto al cambio de escena en el vídeo, situando la distorsión antes y después del cambio de escena.

Las conclusiones extraídas al respecto tienen implicaciones directas en codificación de vídeo, pues, el primer plano después de un cambio de escena se suele codificar como Intra, que, en caso de existir numerosos cambios de escena en una secuencia, daría lugar a la degradación de calidad del resto de planos o al descarte de los mismos para ajustarse al presupuesto de bits asignado. Por lo tanto, aplicando las propiedades de enmascaramiento demostradas en [28] se podría asignar una QP adecuada o descartar planos próximos al cambio de escena sin afectar a la calidad del vídeo.

2.2.4.3 Enmascaramiento por movimiento

La codificación de vídeo con consideraciones perceptuales requiere del conocimiento del foco de interés visual del observador en la escena para completar el conjunto de herramientas a utilizar en su beneficio junto con las propiedades del HVS. Una vez que el interés del observador es conocido, se puede dar prioridad a ciertas zonas del plano de una secuencia para conservar el detalle que contiene, mientras se distorsionan regiones de menor interés visual. El movimiento de áreas correspondientes a un objeto concreto en la escena capta la atención del usuario; si el movimiento es además destacable el observador tiende a realizar un seguimiento de dicho objeto.

En relación con la atención visual del observador, es necesario considerar el movimiento de los ojos y para ello, en primer lugar, la motivación de dicho

movimiento. Realizando un seguimiento del movimiento, podemos llegar a conocer aquello considerado interesante por el observador y, en definitiva, lo que capta su atención en una escena determinada, información de gran importancia en algoritmos de compresión de vídeo, que asignarán mayor tasa de bits a aquellas consideradas de interés subjetivo.

El fenómeno de la atención visual comenzó a estudiarse hace alrededor de un siglo, mediante observaciones oculares e introspecciones, tecnológicamente limitadas. [19] describe la visión como un proceso cíclico compuesto por los siguientes pasos.

(1). Dada una imagen como estímulo, la escena completa es vista primero en paralelo a través de visión periférica y a baja resolución. En este contexto, las zonas interesantes parecen sobresalir del plano, captando la atención del observador por su situación en la escena, para más tarde inspeccionarlas detalladamente.
(2). Una vez localizada la zona de interés, los ojos son rápidamente situados en la misma.
(3). Cuando el movimiento de los ojos finaliza, la fóvea está situada en la región de interés, de modo que las características de esa zona se observan de forma detallada a alta resolución.

A este ciclo se le considera modelo *bottom-up*. Ha sido una de las bases poderosas de los modelos computacionales de investigación visual, pero es un modelo incompleto, pues deja cuestiones en el aire como qué tipo de características llaman la atención del observador, para lo que se recurre al HVS e investiga qué regiones del cerebro son las responsables de la repuesta y la interpretación del estímulo visual que han capturado los ojos; así, se ha demostrado que el HVS responde enérgicamente ante ciertos estímulos como los bordes, y de manera débil a áreas homogéneas por ejemplo.

Otra cuestión a considerar sería que la atención no siempre está asociada con la porción de la escena contenida en la fóvea, pues podemos voluntariamente disociar la atención de esta región, lo que supone un problema común en métodos de seguimiento de ojos pues sólo se pueden seguir los movimientos abiertos de los ojos, no los encubiertos de atención visual.

Existe un gran número de trabajos de investigación que tienen en cuenta este modelo de atención visual, en concreto, en el campo de la codificación de

vídeo, se han desarrollado técnicas de *eye-tracking* eficaces para seguimiento de objetos en movimiento, con el objetivo de encontrar zonas vulnerables a la distorsión. Una muestra de este tipo de modelos podría ser el presentado en [1]; en él se menciona el proceso *bottom-up* correspondiente al comportamiento de la atención humana desarrollado anteriormente, y además, el artículo incluye un proceso adicional denominado *top-down*, que, a diferencia del *bottom-up*, es controlado intencionadamente por el cerebro para dirigir la atención con el fin de realizar una tarea; un modelo computacional que simula este proceso se puede encontrar en [22].

A continuación, se destacan una serie de trabajos de codificación de vídeo donde se proponen o recurren a modelos y medidas de cantidad/complejidad de movimiento que tienen como objetivo detectar zonas enmascarables o simplemente pretenden añadir una mejora en algún sistema diseñado previamente.

En primer lugar, [1] incluye un modelo de atención de movimiento que junto con un modelo de texturas, forman parte de un sistema de asignación de bits propuesto basado en el concepto de sensibilidad a la distorsión visual (VDS, *Visual Distortion Sensitivity*). Este artículo persigue conseguir una segmentación simple de los objetos significativos del plano a partir de la estimación de movimiento global o de cámara, y para ello, elabora un modelo de sensibilidad a la distorsión visual efectivo para indicar regiones perceptualmente importantes, en concreto, las zonas más afectadas serán aquellas que presenten cierto nivel de movimiento y textura aleatoria. El modelo de atención de movimiento presentado es de baja complejidad computacional, pues se compone del cálculo de tres inductores: uno de intensidad, otro de coherencia espacial y otro de coherencia temporal. El primero de ellos, el inductor de intensidad, se corresponde con la intensidad de movimiento de un MB basada en el módulo del vector de movimiento correspondiente a dicho MB. Por su parte, los inductores de coherencia espacial y temporal se basan en el concepto de entropía del vector de movimiento. La coherencia espacial mide el grado de coherencia de los vectores de movimiento que componen un objeto. La coherencia temporal mide la correlación temporal de los mismos, de modo que si un objeto se mueve rápido, poca coherencia temporal presentará y, por tanto, mayor atención al mismo habrá por parte del ojo humano. Finalmente, el índice de atención de movimiento de un MB es el resultado del producto de los inductores anteriormente calculados:

$$MI_{nij} = I_{nij} \cdot Ct_{nij} \cdot \left(1 - I_{nij} \cdot Cs_{nij}\right) \quad (5)$$

Una vez calculados los índices característicos de cada modelo, del movimiento y las texturas, el índice de sensibilidad a la distorsión (VDSI) se calcula en función de ambos. A su vez, el valor final del escalón de cuantificación asignado a cada MB se calcula en función del VDSI.

Del mismo modo, el trabajo [17] destaca por un analizador de texturas y otro de movimiento en su algoritmo propuesto. En concreto, señala la importancia de la detección de movimiento en una escena; indica que el movimiento y los objetos en primer plano tienen una fuerte influencia en el observador, y también, que el usuario da prioridad a objetos que tienen un movimiento notable. Por lo tanto, todas estas asunciones sobre atención al movimiento se tienen en cuenta en la codificación del siguiente modo: mediante un análisis *foreground-background*, es decir, se busca el primer plano y el fondo de cada plano. Se recurre a un método de compensación de movimiento global-local encargado de analizar el movimiento de la escena, obteniendo una clasificación de movimiento; se supone un modelo de perspectiva de 8 parámetros que establece panorámica, zoom y movimiento de rotación 3D. Adicionalmente, en este trabajo se trata la inconsistencia temporal, un problema común en codificación de vídeo que da lugar a distorsiones visibles en el resultado final. En este modelo concreto se produce al aplicar cada analizador plano a plano, afectando a la calidad visual del vídeo reconstruido, pues las regiones donde la distorsión es aplicada de un plano al siguiente pueden variar considerablemente, resultando molesto; como solución a este problema, proponen utilizar el modelo de movimiento para delimitar las regiones de detalles irrelevantes (del fondo) a "deformar" plano a plano.

Por otro lado, otro modo de obtener una medida de cantidad de movimiento sin aplicar estimación de movimiento, y por tanto, no tener ningún tipo de información sobre los vectores asociados, puede estar basada en la cantidad de bits asignados por plano. La referencia [2] determina una medida de complejidad de movimiento que depende de los bits asignados a los planos ya codificados; es decir, como antes de la codificación entrópica no podemos conocer los bits que van a ser asignados al plano actual, se recurre a una estimación de dicha cantidad de bits, que recurre a la correlación temporal entre planos consecutivos. Sin embargo, esta medida no influye sobre el cálculo de la QP asignada a cada MB, como sucedía en el caso anterior, sino que se tiene en cuenta en el cálculo de la tasa de bits a asignar a cada plano.

En el siguiente trabajo referenciado, [16], no se elabora ninguna medida de complejidad de movimiento, ni se aplica ninguna técnica de extracción de características del movimiento de una escena de carácter subjetivo, sino que se presenta una mejora del cálculo del mapa de vectores de movimiento para el codificador H.264/AVC, basándose en bloques Inter 8x8. El método, detallado en [16], propone tomar la decisión de estimación de movimiento de un bloque considerando tanto a bloques adyacentes pasados como futuros, según indica la Figura 2.15.

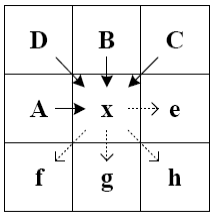


Figura 2.15. Localizaciones de MBs vecinos.

El método propuesto se denomina JME (*Joint Motion Estimation*), y busca para las cuatro particiones de un MB los vectores de movimiento correspondientes de forma conjunta, es decir, las estimaciones de cada partición se combinan mediante bucles anidados, aumentando la complejidad computacional con respecto al método tradicional. La función de coste que se pretende minimizar se basa en la SAD (*Sum of Absolute Differences*), que consiste en la suma de las diferencias absolutas entre el bloque actual y su bloque de predicción, píxel a píxel. Para agilizar los cálculos del nuevo método, se propone precalcular las SAD de cada partición de bloque candidata. Añadido este método propuesto en un codificador característico del estándar, se observa en los resultados de las pruebas experimentales llevadas a cabo una mejora perceptible en el campo de vectores obtenido con respecto a la técnica original.

2.2.4.4 Enmascaramiento por excentricidad

Es necesario definir el concepto de excentricidad para facilitar la comprensión de la base de este tipo de enmascaramiento. En el globo ocular, la excentricidad es la distancia existente desde cualquier punto de la retina al centro de la misma. La resolución de visualización en la retina varía con la excentricidad (véase Figura 2.16) pues, como ya se mencionó en apartados anteriores, los conos

se ubican principalmente en la zona de excentricidad 0º, punto denominado centro retiniano o fóvea, y su densidad decrece con la excentricidad.

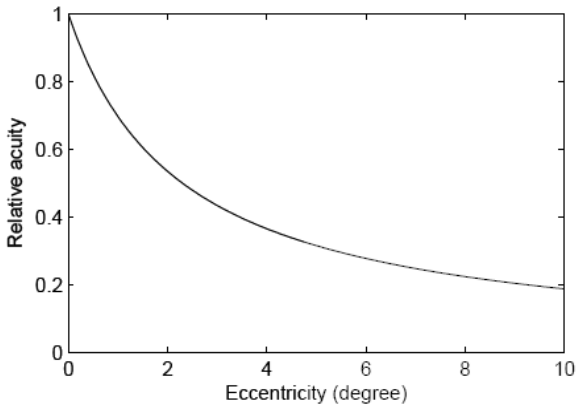


Figura 2.16. Agudeza visual.

Estudios realizados en el ámbito de la codificación de vídeo que se basan en esta característica del HVS, consiguen determinar en cada plano de una secuencia de vídeo la región visualizada a la mayor resolución por el observador, y por tanto, considerada de interés subjetivo elevado; este área es conservado en la medida de lo posible, mientras el resto del plano sufre una degradación a determinar. Por lo tanto, el conocimiento de esta información correspondiente a la fóvea permite explotar la redundancia perceptual global.

Un ejemplo de este tipo de enmascaramiento es [12], que presenta un modelo JND basado en la fóvea para conseguir una compresión eficiente que no degrade la calidad subjetiva del vídeo codificado resultante. Recurre a modelos de JND espacial y temporal, que se basan en que la redundancia perceptual en el dominio espacial se debe a la sensibilidad del HVS al contraste de luminancia y los efectos de enmascaramiento y en que el enmascaramiento temporal puede ser representado por el promedio entre la luminancia de planos consecutivos.

Por otro lado, el artículo [3] presenta un modelo de asignación de bits basado en consideraciones perceptuales en el que el presupuesto asignado al fondo disminuye en función de la distancia al primer plano. Para ello, recurre a un factor de sensibilidad visual que pondera el presupuesto de bits y que además decrece exponencialmente según la distancia absoluta del MB en estudio a la región más cercana perteneciente al primer plano. De esta manera, se da prioridad a regiones de mayor interés subjetivo, asignándoles una calidad mayor, mientras

que el resto, a medida que se alejan del centro de interés la degradación de calidad que sufren aumenta.

2.2.4.5 Detección de la región de interés

Una vez descritos todos los tipos de enmascaramientos visuales disponibles en codificación perceptual de vídeo es necesario conocer otra vertiente destacada. Se trata de la localización de las regiones de interés o ROI de una secuencia.

Las ROI son aquellas áreas de las que está pendiente el ojo humano bien en una imagen estática, o en una secuencia de imágenes; estas zonas tienen que ser consideradas para codificarlas convenientemente.

A diferencia del enmascaramiento visual, que busca regiones de interés donde la distorsión es menos perceptible por el ojo humano basándose en su textura, forma y/o movimiento, la detección de ROI consiste generalmente en estimar un mapa de importancia de los elementos de la escena, por lo tanto, se requieren métodos de segmentación para obtenerlo. Este método suele estar destinado a escenarios muy concretos donde las regiones de interés están bien definidas, como por ejemplo, secuencias pertenecientes a videoconferencias donde la región de interés es el rostro de la persona que habla, o correspondientes a deportes, donde el centro de atención suelen ser los jugadores. Por lo tanto, su uso generalizado no parece recomendable, puesto que enfrentarse a un vídeo de mayor complejidad supondría una técnica de detección complicada.

A continuación se incluyen algunos ejemplos acerca de métodos desarrollados sobre detección de ROI recogidos en la bibliografía. En primer lugar, [3] presenta una propuesta de asignación de bits objetivo a cada plano en la que, dado un presupuesto de bits, el parámetro de cuantificación para el *primer plano* o *foreground* se optimiza para encontrar una calidad objetivo; después, los bits restantes se asignan al *fondo* o *background* de modo que la calidad del mismo disminuye gradualmente con la distancia al *foreground*.

Por otro lado, [5] realiza una segmentación burda del plano en zonas con movimiento y zonas estáticas, para reducir la complejidad computacional que el proceso pudiera suponer. Señala que debido a la sensibilidad del HVS a regiones en movimiento, es razonable sacrificar la calidad perceptual de las regiones sin

movimiento a favor de aquellas que sí lo presentan; por lo tanto, más bits se asignan a estas regiones.

Para detectar las zonas con movimiento recurre a la diferencia entre planos; se establece un umbral de tasa de movimiento que si es superado por la tasa correspondiente de cada MB, éste será considerado ROI, sino, se determina non-ROI.

Otra forma de detectar regiones de interés es propuesta en [7]; se basa en la no uniformidad de la distribución de los fotorreceptores en la retina, que, como fue mencionado en apartados anteriores, se concentra en la fóvea; por lo tanto, esto fundamenta la codificación de una imagen sin calidad uniforme, pues hay regiones que toleran una distorsión mayor.

Los métodos presentados son: uno interactivo y otro de propósito general. Por un lado, el primer método es de tipo *eye-tracking* que realiza un seguimiento del ojo del observador, determinando las zonas más destacadas para el usuario, para, posteriormente, codificarlas con mayor fidelidad, mientras el resto del plano queda degradado. Los resultados obtenidos son eficientes, pero como está limitado a un único observador, se recurre al segundo método. La otra propuesta establece un mapa de ROI teniendo en cuenta características visuales de cada plano como el contraste, la intensidad, orientación, etc.

Para finalizar con este apartado sobre la detección de regiones de interés en los planos de una secuencia es necesario señalar el trabajo realizado en [13], donde se comparan dos experimentos para modelar la atención visual: *ROI selectiva* frente a *Fijación Visual de Patrones* (*VFP*). Con respecto a la ROI selectiva, el experimento es realizado por 30 observadores, cuya función es seleccionar regiones rectangulares de tamaño indeterminado en un conjunto de imágenes. Por su parte, el VFP consiste en un experimento *eye-tracking* realizado a 15 usuarios, donde se les muestra el mismo conjunto de imágenes expuesto en la prueba anterior y adicionalmente versiones distorsionadas de ellas; los observadores deben evaluar la calidad. De los experimentos se concluye que las caras humanas y en concreto los ojos de las mismas son de gran interés visual, y que el VFP es más preciso al reflejar la atención visual pues en él existe un factor temporal que no puede ser capturado por la ROI.

2.2.5 Métodos de inclusión de distorsión localizada

2.2.5.1 Incremento del parámetro de cuantificación

Se trata de un método de introducción localizada de distorsión recurrente en codificación de vídeo, que proporciona una mayor compresión y busca conservar la calidad subjetiva de los vídeos codificados. En la bibliografía existen varios artículos que incluyen una técnica de asignación de la QP adaptativa a una medida característica del grado de distorsión o de enmascarabilidad soportado por el ojo humano.

En esta técnica, una vez se realiza un análisis de regiones enmascarables en los planos de la secuencia a estudiar, la degradación de calidad introducida en ellas se puede conseguir mediante el ajuste de la QP según el grado de enmascarabilidad que presente la zona. De esta manera mayores escalones de cuantificación se utilizarán en regiones de baja importancia subjetiva, mientras el valor de la QP se disminuye para las regiones de mayor interés subjetivo, con el fin de preservar su detalle.

La primera referencia que presenta un sistema de cálculo de QP según las regiones enmascarables es [1]. Como ya se ha mencionado con anterioridad, este trabajo introduce el concepto de Índice de Sensibilidad a la Distorsión Visual (VDSI), es decir, capacidad de la visión humana para detectar distorsión en secuencias de vídeo. Esta magnitud se obtiene a partir de dos medidas que caracterizan el movimiento y las texturas de la secuencia, ya detalladas en el apartado 2.2.4, referente a los tipos de enmascaramiento. En definitiva, la expresión que calcula el parámetro de cuantificación es la siguiente:

$$QP'_{nij} = QP_{nij} + \left(1 - \frac{VDSI_{nij}}{VDSI_{max}}\right) \cdot \Delta Q \qquad (6)$$

Donde QP_{nij} es el parámetro de cuantificación a nivel de plano que asigna el algoritmo de control de tasa del codificador correspondiente, y ΔQ es el parámetro para limitar el valor que va a modificar la QP_{nij} y cumple $\Delta Q \geq 0$.

La asignación de QP también puede depender del cálculo de la denominada JND que trata [12]. El modelo JND desarrollado en este trabajo se basa en las características de contraste y efectos de enmascaramiento de la secuencia de estudio correspondiente, por ello, elabora un modelo espacial-temporal basado en

la fóvea (FJND, *Foveated JND*), con el objetivo de conseguir una representación de la escena visual más inteligente. En relación a la variación del parámetro de cuantificación, éste se asigna a cada MB de forma independiente; su magnitud depende del escalón de cuantificación correspondiente al plano (Q_r, determinado por el control de tasa del codificador), y del peso de distorsión notable (w_i), calculado a partir del promedio FJND característico del MB (s_i). Se observa en la expresión (7) que se realiza el cociente entre estos dos parámetros para obtener la QP característica de cada macrobloque y, según la definición de w_i, la relación entre la FJND del MB con respecto al promedio FJND del plano (s) determina el valor definitivo.

$$Q_i = Q_r \Big/ \sqrt{w_i} \qquad (7)$$

$$w_i = a + b\frac{1 + m\cdot exp(-c\cdot\frac{s_i - s}{s})}{1 + n\cdot exp(-c\cdot\frac{s_i - \bar{s}}{s})} \qquad (8)$$

Por otro lado, en lugar de obtener la magnitud del parámetro de cuantificación basándose en una medida de sensibilidad a la distorsión, [3] parte del presupuesto de bits necesarios para codificar cada macrobloque. En primer lugar, se predice el presupuesto de bits para el *foreground* o *primer plano* (P_{FG}), y luego se asignan los bits restantes al fondo (*background*), de manera que la calidad del mismo disminuye gradualmente con la distancia al primer plano. La degradación de calidad del fondo está controlada por el factor de sensibilidad visual (S), que disminuye con la distancia y está ponderado por el factor α (véase (9)), inicializado de manera aleatoria para el primer plano, y al valor del plano anterior para el resto con el fin de conseguir una cierta consistencia.

$$S = e^{-d/\alpha} \qquad (9)$$

$$P_{BG} = P_{FG}\cdot S \qquad (10)$$

Por lo tanto, la QP es dependiente de la distorsión del fondo (D_{BG}), la desviación estándar del MB actual (σ_i) con respecto del resto de MB, y E, la relación entre la verdadera distorsión y la calculada por el modelo.

$$Q_i = \sqrt{\frac{12\cdot N_b\cdot D_{BG}\cdot\sigma_i}{E\cdot\sum_{i=1}^{N_b}\sigma_i}} \qquad (11)$$

Para finalizar, señalar que el estándar H.264/AVC permite la posibilidad de modificar de forma adaptativa el tamaño del escalón de cuantificación y la zona "muerta" (*deadzone*) o de cuantificación cero para mejorar la eficiencia de codificación. Se realiza mediante el "*offset*" de redondeo, "*s*".

$$Z = \frac{|X|}{QP} + s \cdot sgn(X) \qquad (12)$$

$$X' = Z \cdot QP \qquad (13)$$

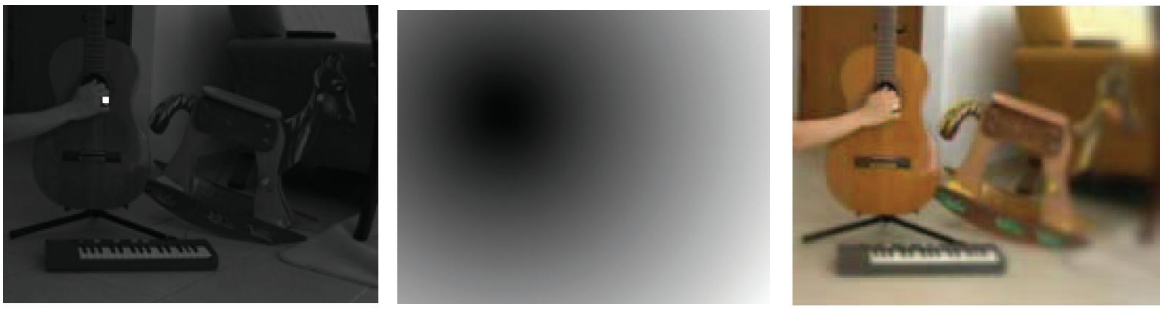
Este parámetro tiene la ventaja de regular el proceso de cuantificación sin la necesidad de transmitir parámetros adicionales al decodificador (véase (13)). Dado el paso de cuantificación *q*, la *zona "muerta"* aumenta cuando *s* decrece y más coeficientes transformados serán cuantificados a cero, resultando una tasa de bits más baja. En [35] se presenta un nuevo algoritmo de control de tasa con *offsets* de redondeo adaptativos, denotados como ARO (*Adaptive Rounding Offsets*), que ajustan tanto la *QP* como la *s* para conseguir eficacia en tasas altas a nivel de plano. Este mecanismo podría aplicarse para mejorar la eficiencia de codificación en regiones de interés, previamente delimitadas según las consideraciones del HVS pertinentes.

2.2.5.2 Técnicas de "blurring"

El término *blurring* se denomina al efecto de desenfoque o de falta de definición presente en un plano. Consiste en la ausencia de detalle espacial en imágenes de actividad alta/moderada, como pueden ser aquellas con texturas bien definidas o bordes pronunciados.

Este tipo de distorsión aparece por diversas causas. Para MBs Intra, el efecto *blurring* está relacionado con la supresión de coeficientes de alta frecuencia de la DCT después de la cuantificación, quedando solo los coeficientes de más baja frecuencia como representantes del contenido del bloque; este hecho puede coincidir con la aparición de los efectos de bloques y de mosaico. Mientras tanto, para MBs tipo P, surge como consecuencia del uso de MB de referencia con falta de detalle espacial. Y, para MBs tipo B, el *blurring* aparece al interpolar las predicciones *backward* y *forward*, que tiene como resultado el promedio de los contenidos de la predicción bidireccional final.

Por otro lado, se han desarrollado técnicas que recurren a este tipo de distorsión, pues, introducida de forma localizada en el proceso de codificación de vídeo se pueden conseguir resultados eficientes. Una muestra de ello es el trabajo realizado en [24]; presenta una nueva técnica de codificación basada en atención no sólo visual, sino también acústica. En concreto, recurriendo a una técnica CCA (*Canonical Correlation Analysis*) se localiza la región del plano asociada a la fuente acústica de la secuencia (según indica el punto blanco situado en la Figura 2.17.(a)); una vez obtenida, se produce un mapa de prioridad que asigna pesos a las distancias entre cada píxel y la energía localizada más cercana, de modo que el mayor valor de prioridad se corresponde con la región localizada y disminuye a medida que la distancia a dicho punto aumenta (véase Figura 2.17.(b)). El efecto *blurring* se aplica sobre el mapa de prioridad calculado y consiste en una pirámide de filtros paso bajo Gaussianos con L niveles; las zonas de baja prioridad se relacionan con niveles bajos de *blurring* y el nivel más alto con la mayor prioridad.



(a) (b) (c)

Figura 2.17. Ejemplo del método propuesto. (a) Localización fuente acústica. (b) Mapa de prioridad. (c) Imagen con *blurring* para L=6.

2.3 Medidas de calidad subjetiva

En este apartado se realiza una breve introducción sobre los modelos de medida de calidad de vídeo e imagen existentes, haciendo hincapié en aquellos a los que se recurre a lo largo de este trabajo. La referencia [33] recoge un estudio detallado sobre los modelos de medida de calidad de vídeo, por ello se recurre a ella a lo largo de los siguientes sub-apartados.

En primer lugar, es necesario conocer el concepto de calidad para el observador antes de desarrollar cualquier medida de calidad subjetiva. Consiste, por tanto, en aquellas propiedades inherentes relativas a un objeto original que

una secuencia de vídeo es capaz de representar, es decir, la exactitud entre ambos. Así, en los modelos de medida de calidad, se recurre a la comparativa de dos secuencias: la original o de referencia, y la secuencia que ha sufrido algún tipo de distorsión.

Por otro lado, existen dos grupos principales de modelos de medida: métricas de calidad de imagen subjetivas (requieren observadores) y objetivas (medidas matemáticas). A su vez, las objetivas, en función de la incorporación o no del HVS, se dividen en: modelos de *detección de error*, basados en *distorsión estructural* e *híbridos*.

El uso de un modelo de calidad u otro, viene dado por las propiedades que caractericen a cada uno, como la velocidad de cómputo y la complejidad, la precisión de sus resultados o la robustez.

A continuación, se detallan las medidas objetivas principalmente empleadas durante muchos años, así como las principales medidas basadas en el sistema visual humano más recientes.

2.3.1 MOS (Mean Opinion Score)

Antes de presentar los modelos de medida de calidad más destacados es necesario definir el concepto de MOS. En codificación de vídeo y audio se recurre a esta medida de calidad para disponer de un indicador numérico característico de la calidad percibida después de la compresión y/o la transmisión.

Como su propio nombre indica se trata de la media aritmética de todas las puntuaciones de opinión subjetiva. Las puntuaciones se mueven entre los valores 1 y 5, siendo el 1 la peor puntuación de calidad y 5 la mejor de ellas, según se indica a continuación.

MOS	CALIDAD
5	EXCELENTE
4	BUENA
3	ACEPTABLE
2	MEDIOCRE
1	INACEPTABLE

Tabla 2.1. Puntuaciones MOS.

En experimentos de evaluación de calidad subjetiva, los individuos sometidos a dicha prueba evalúan las secuencias observadas asignándolas una puntuación según la correspondencia que recoge la Tabla 2.1.

2.3.2 Medidas objetivas basadas en distorsión comparativa

Este tipo de medidas matemáticas asumen que la pérdida de calidad está directamente relacionada con la potencia de la señal de error, de modo que para evaluar la calidad de una imagen se cuantifica el error entre la señal distorsionada y la original.

El principal inconveniente que presentan es que sólo son eficientes cuando los errores se corresponden con ruido adicional no correlacionado con la señal. A continuación, se listan las medidas más conocidas, de entre las cuales, destaca el uso del error cuadrático medio o *MSE* y la relación señal a ruido de pico, *PSNR*.

- *Mean Squared Error*, MSE.

$$MSE = \frac{\sum_{i=0}^{N}\sum_{j=0}^{M}(x(i,j) - d(i,j))^2}{NM} = \frac{(x - \mathbf{d})(x - \mathbf{d})^T}{NM} \qquad (14)$$

Donde N, M son las dimensiones vertical y horizontal de la imagen en píxeles; $x(i,j)$, $d(i,j)$ imagen original y distorsionada respectivamente en notación matricial; y, $\mathbf{x}$ y $\mathbf{d}$ imagen original y distorsionada respectivamente en notación lexicográfica.

- *Peak Signal to Noise Ratio*, PSNR.

$$PSNR(dB) = 10 \cdot \log_{10} \frac{L^2}{MSE} \qquad (15)$$

Donde L es el rango dinámico de los valores de los píxeles. Para una señal monocromática de 8 bits/píxel, L es igual a 255.

- *Root Mean Squared Error*, RMSE.
- *Normalized Mean Squared Error*, NMSE.

- *Signal to Noise Ratio*, SNR.
- *Signal to Error Ratio*, SER.
- *Standard Deviation*, STD.

2.3.3 Medidas basadas en detección de error

Se trata del primer conjunto de medidas de calidad basadas en el sistema visual humano, pues se basan en la capacidad de éste en detectar errores en imagen o vídeo. Así, el principio básico es considerar la secuencia distorsionada como suma de una señal de referencia de calidad perfecta y una señal de error; por tanto, el algoritmo de evaluación de calidad debe determinar la potencia de la señal de error y analizar en qué medida afecta a la percepción humana, según las características del HVS.

El esquema general de un algoritmo de medida de calidad basada en detección de error se corresponde con la siguiente figura.

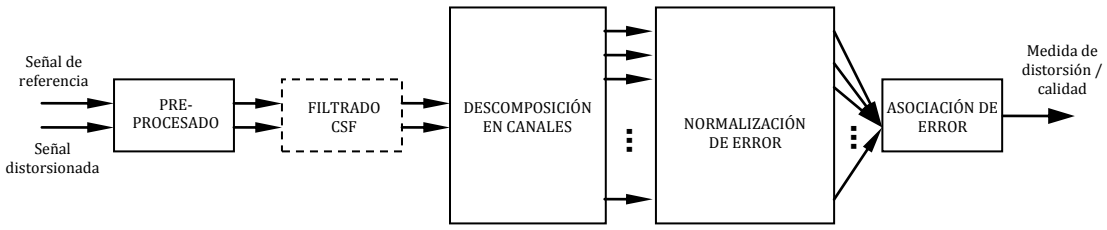


Figura 2.18. Esquema de un sistema de medida de calidad basado en detección de error.

Por lo general, la primera etapa del algoritmo suele consistir en un pre-procesado de las señales de entrada. Se compone de una alienación espacial y/o temporal de las señales de referencia y distorsionada, seguida de una transformación del espacio de color para que se adapte mejor al HVS; posteriormente, otras de las operaciones posibles son la calibración de los dispositivos de visualización, un filtrado paso bajo para simular la PSF (*Point Spread Function*) o función de dispersión de punto característica del ojo. En el caso de vídeo, las señales de referencia y distorsionada necesitan ser convertidas a su correspondiente estímulo de contraste para simular la adaptación a la luz.

El filtrado CSF, según la sensibilidad al contraste, presente en el diagrama puede realizarse antes de la descomposición de canal o bien, posteriormente, asignando a cada canal un factor de ponderación en función de la sensibilidad a la

correspondiente frecuencia. Esta descomposición de canales permite dividir el estímulo visual en diferentes subbandas espaciales y temporales; para ello, se recurre a transformaciones sencillas como la transformada Wavelet o la DCT, u otras más sofisticadas como funciones 2D de Gabor que representan las zonas corticales encargadas de la visión, o la *cortex transform* (transformación de corteza) de Watson [32].

Después, la normalización del error y el enmascaramiento se aplican sobre cada canal. En el ajuste del umbral de visibilidad intervienen la señal de error entre las dos señales de entrada, la señal de referencia y la sensibilidad del HVS para el canal en ausencia de los efectos de enmascaramiento (conocida como sensibilidad base). El umbral de visibilidad se utiliza luego para normalizar la señal de error. Esta normalización típicamente convierte el error en unidades JND, es decir, diferencia apreciable, donde un JND de 1.0 indica que la distorsión en ese punto, en ese canal está justo en el umbral de visibilidad.

Para finalizar, es necesario asociar las señales de error de los diferentes canales. La mayoría de los métodos de evaluación recurren a la asociación de error Minkowski (véase (16)).

$$E = \left(\sum_l \sum_k \left|e_{l,k}\right|^{\beta}\right)^{1/\beta} \quad , \qquad (16)$$

donde $e_{l,k}$ es el error enmascarado y normalizado del coeficiente k-ésimo en el canal l-ésimo, y β es una constante con un valor normalmente entre 1 y 4.

Una vez descritas las partes que componen un algoritmo de evaluación de calidad basado en detección de error es necesario señalar las limitaciones características. Pues, se intenta simular la calidad perceptual pero el HVS es muy complejo, un sistema altamente no lineal del cual no se tienen los suficientes conocimientos. Por tanto, estos algoritmos se basan en ciertas suposiciones que dan lugar a sus deficiencias, una de ellas, que da lugar a la aparición del siguiente tipo de medidas es el hecho de que la descomposición de canal efectivamente elimina correlaciones en la estructura de la imagen, de manera que la potencia de la señal se determina por las magnitudes de los coeficientes, de modo que la información estructural se pierde, y, como consecuencia, dos imágenes distorsionadas de manera diferente basadas en una misma imagen original pueden generar señales de error absoluto idénticas; así, [33] demuestra la ineficiencia del error Minkowski en ciertas situaciones.

2.3.4 Medidas basadas en distorsión estructural

Como se ha citado en el apartado anterior, las medidas de calidad basadas en detección de error consideran cualquier tipo de distorsión como cierto tipo de error. Sin embargo, diferentes estructuras de error tendrán efectos distintos en la calidad percibida. Pues, los métodos de descomposición lineal del canal no pueden eliminar completamente correlaciones en las estructuras de la señal, y, por tanto, la medida de error de Minkowski no puede capturar estas correlaciones estructurales.

Surge una filosofía alternativa acerca de la medida de calidad de imagen y vídeo, que, sustituye la medida del error por la de distorsión estructural. Se basa en que el HVS extrae información estructural del campo de visión, de modo que la medida de la distorsión estructural podría ser una buena aproximación de la distorsión percibida (según Wang, Bovik y Lu en 2002). Así, distorsiones estructurales significativamente diferentes pueden caracterizarse por la misma cantidad de error, mientras la calidad percibida de cada una es muy distinta. Una observación importante de este nuevo tipo de medidas es que consideran la degradación de la imagen como pérdida de información estructural percibida; de este modo, la expansión de contraste, por ejemplo, mejora la calidad porque la información estructural de la imagen se preserva, sin embargo, en términos de error, existe una gran diferencia con la imagen original.

Por el momento, existen dos formas de implementar algoritmos de medida de calidad basados en esta nueva filosofía. El primero de ellos consiste en desarrollar un entorno de descripción de características de las imágenes naturales, que cubra la mayoría de la información estructural de una imagen. En ese entorno, los cambios en la información estructural entre la señal original y la distorsionada pueden ser cuantificados mediante un índice de calidad, como por ejemplo, el propuesto por Wang y Bovik, modelado como combinación de tres factores: pérdida de correlación, distorsión de media y distorsión de contraste. La segunda es diseñar un método de comparación de estructuras que pueda comparar la similitud o la diferencia estructural entre las imágenes de forma directa.

2.3.5 Medida de Calidad de Vídeo VSSIM

El algoritmo VSSIM o Índice de Similitud Estructural de Vídeo, hace uso del Índice SSIM (*Structural SIMilarity index*) de imágenes estáticas. Esta medida, basada en distorsión estructural, será la base fundamental del método utilizado en la evaluación subjetiva ([33]) del algoritmo propuesto. El índice SSIM consiste en una aproximación de medida que separa la luminancia, el contraste y las distorsiones estructurales, a partir de la expresión (17).

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \qquad (17)$$

C_1 y C_2 son dos constantes: $C_1 = (K_1 \cdot L)^2$ y $C_2 = (K_2 \cdot L)^2$, donde L es el rango dinámico de los valores de los píxeles (para 8 bits/píxel, L=255); por su parte, K_1 y K_2 son dos constantes a fijar experimentalmente a 0.01 y 0.03 respectivamente.

La Figura 2.19 se corresponde con el diagrama del sistema de evaluación de calidad de vídeo VSSIM, que, como se puede observar, se mide en tres niveles: área local, fotograma y secuencia.

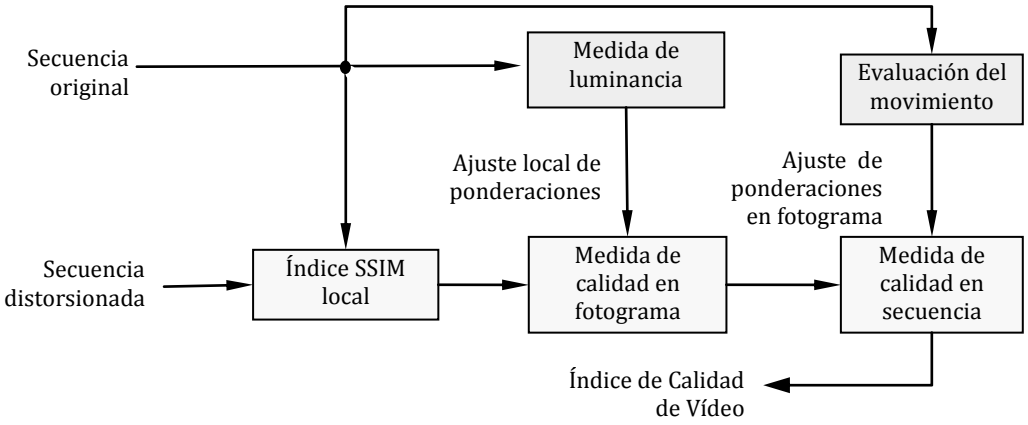


Figura 2.19. Diagrama de bloques del algoritmo VSSIM.

El primer paso a realizar consiste en el cálculo del índice SSIM de ventanas de muestreo aleatorias de tamaño 8x8 píxeles, tanto de un plano de la secuencia original como de la distorsionada. El índice SSIM se aplica sobre cada componente Y, Cb, Cr de forma independiente, que, combinados mediante una suma ponderada generan un índice de calidad local.

Posteriormente, en el segundo nivel de evaluación, los valores de calidad local se combinan en un índice a nivel de fotograma, donde cada ventana tiene asociado un valor de ponderación. El ajuste de estas ponderaciones se basa en la consideración de que las zonas oscuras normalmente no atraen la atención del observador, por lo que se deben ponderar con pesos menores; para ello, recurre al valor medio de la componente Y como estimador de luminancia local.

Finalmente, el tercer nivel de calidad es el correspondiente al de la secuencia de vídeo completa. En el cálculo del índice de calidad global se realiza también un ajuste de los pesos que ponderan el índice a nivel de fotograma. Para ello se recurre a un algoritmo de estimación de movimiento basado en bloques para evaluar el movimiento con respecto al fotograma siguiente. Un desenfoque aplicado sobre una secuencia con movimiento rápido no es tan importante a nivel perceptual porque también tiene lugar un importante desenfoque perceptual causado por el movimiento, mientras un desenfoque muy alto sobre un vídeo con un movimiento lento puede provocar distorsiones desagradables perceptibles.

2.3.6 Índice MOVIE

Para finalizar con las medidas de calidad subjetiva, se presenta el índice MOVIE (*MOtion-based Vídeo Integrity Evaluation*), también utilizado en la evaluación del sistema propuesto.

Lo más destacado de esta medida es que integra tanto medidas de distorsión espaciales como temporales y utiliza una localización espacio-temporal, descomposición multiescala de los vídeos de referencia y de test mediante un conjunto de filtros espacio-temporales de Gabor, pues un proceso multiescala se aproxima más al proceso de percepción humana. El índice MOVIE tiene dos componentes: índice MOVIE Espacial e índice MOVIE Temporal. El cálculo de cada uno de desarrolla con detalle en [34], aunque, a continuación, se incluyen algunas pautas acerca del algoritmo presentes en esta referencia.

Así, el primer paso consiste en la descomposición lineal mediante filtros separables de Gabor que tienen la misma desviación estándar a lo largo de las coordenadas de frecuencia espacial y temporal. La implementación realizada en [34] utiliza tres escalas de 35 filtros cada una, añadiendo una contribución adicional al algoritmo base.

Por un lado, el índice MOVIE Espacial recurre a la salida del banco de filtros Gabor pues éstos representan una descomposición de las secuencias de entrada en canales paso-banda. Por tanto, los filtros de Gabor responden individualmente a un rango específico de frecuencias espacio-temporales y orientaciones en el vídeo, y cualquier diferencia en el contenido espectral de la referencia y la secuencia distorsionada es capturada por las salidas Gabor. Así, las distorsiones pueden ser capturadas recurriendo al cálculo del error entre sub-bandas correspondientes de la referencia y la secuencia de test.

Por su parte, el índice MOVIE Temporal permite capturar degradaciones temporales que dan lugar a pérdidas en la calidad del vídeo. La información de movimiento se calcula a partir del vídeo de referencia en forma de campos de flujo óptico, que se obtienen a partir del conjunto de filtros Gabor. Este índice, por tanto, evalúa la calidad del vídeo de test a través de las trayectorias de movimiento del vídeo de referencia.

Finalmente, los índices MOVIE Espacial y Temporal se combinan para obtener una única medida de calidad de vídeo, denominada índice MOVIE, representativa de la secuencia en estudio.

Capítulo 3

Sistema global de enmascaramiento por movimiento

En este capítulo se incluye una descripción general del sistema de enmascaramiento por movimiento que se propone, sin entrar en detalle en los diferentes módulos que lo conforman, con el fin de establecer los objetivos a alcanzar mediante el trabajo realizado.

Una de las propiedades básicas del sistema visual humano es el enmascaramiento. Puede ser ocasionado por diversos factores que caracterizan una escena, como la textura, la respuesta temporal del HVS al movimiento, o la presencia de movimiento perceptible en ciertas regiones de la imagen. El enmascaramiento producido por el movimiento de áreas del plano determinadas de la escena es la limitación que sirve de base en la elaboración del sistema propuesto. En presencia de un movimiento elevado, el ojo humano es incapaz de percibir toda la información (bordes, texturas, contraste,…), por tanto, es menos sensible a posibles distorsiones concentradas en las zonas que presentan dicho movimiento. Además, en una escena con movimiento destacable, éste se considera

una característica significativa que permite realizar una segmentación eficaz que agrupa las regiones del plano en función de la cantidad de movimiento que presentan.

Entonces, en primer lugar, se propone detectar los MBs que presentan mucho movimiento y aplicar sobre ellos una cuantificación más elevada con respecto a la del resto del plano, con el objetivo de conseguir una reducción de la tasa binaria generada tras la codificación sin que ello implique una disminución de la calidad de la secuencia. Para elaborar esta técnica se requiere un clasificador encargado de catalogar cada MB según su cantidad de movimiento: mucho/poco movimiento.

Sin embargo, este clasificador no tiene en cuenta las zonas de interés subjetivo, pues aquellas zonas que se cataloguen como con mucho movimiento podrían no corresponderse con regiones de interés, es decir, en donde el cerebro concentra su mayor atención. En concreto, si diferenciamos entre vídeos con fondo estático y vídeos con fondo en movimiento, las regiones de interés cambian.

Así, en el caso de vídeos con fondo en movimiento, éstos siguen la trayectoria del objeto (u objetos) de interés, por ello, se puede aplicar distorsión a los MBs que pertenecen al fondo, pues presentan movimiento y no forman parte de la región de interés que conviene preservar. Pero, en el caso de vídeos con fondo estático, será considerado de interés todo aquello que presente un movimiento significativo, por lo tanto, en este caso no es conveniente introducir distorsión en zonas con un grado elevado de movimiento, porque es precisamente en ellas donde el usuario centra su atención.

Como consecuencia de todo ello, surge la necesidad de otra clasificación complementaria que se encargue de determinar qué MBs son de interés y cuáles no, para poder asignar un nivel de distorsión adecuado a cada región sin que implique una reducción de la calidad de la secuencia. Entonces, cada MB del plano se caracterizaría de la siguiente manera:

1. Mucho / Poco movimiento
2. Zona de interés (objeto) / Zona de no interés (fondo)

Utilizando esta caracterización se determina que no todas las combinaciones de las salidas de cada clasificador se corresponden con una asignación dura de distorsión, es decir, alta o baja distorsión, debido a su

relevancia subjetiva. Por lo tanto, se pueden aplicar varios niveles de distorsión en la imagen siguiendo este criterio:

	MUCHO MOVIMIENTO	POCO MOVIMIENTO
ZONA DE INTERÉS	DISTORSIÓN MEDIA/BAJA	DISTORSIÓN BAJA
ZONA DE NO INTERÉS	DISTORSIÓN ELEVADA	DISTORSIÓN MEDIA

Tabla 3.1. Asignación de distorsión.

El diagrama de bloques genérico del sistema compuesto por los dos clasificadores citados previamente se corresponde con la Figura 3.1. En la imagen se han ocultado bloques de procesamiento, pertenecientes a este sistema también, que tienen como objetivo mejorar la calidad de la clasificación final; pero, estos bloques añadidos se explican con detalle en el Capítulo 5.

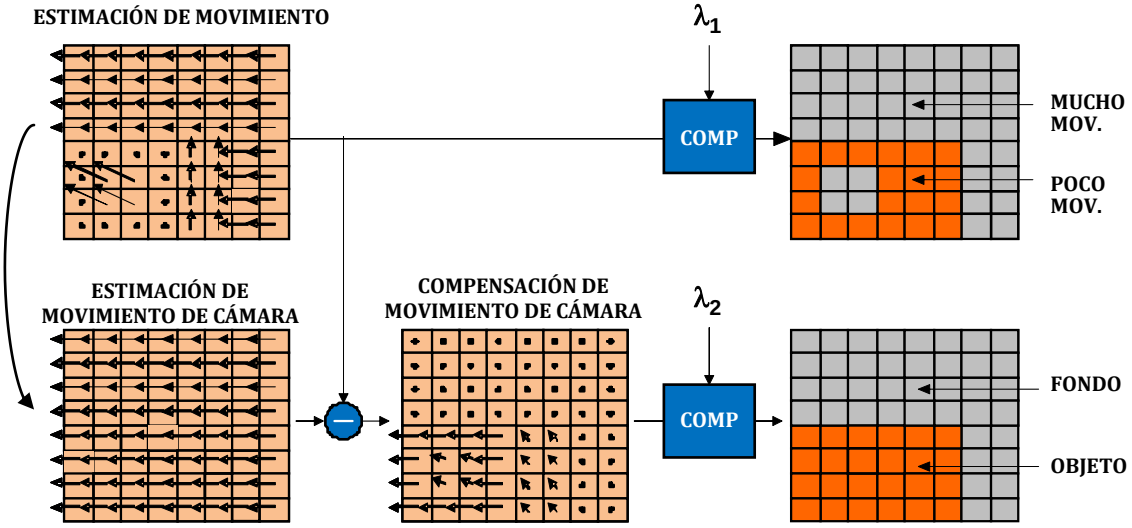


Figura 3.1. Esquema general del algoritmo de enmascaramiento por movimiento.

Como se puede observar, el sistema se divide en dos ramas que parten de un mismo mapa de vectores de movimiento; el mapa, a nivel de MB, representa de forma fidedigna el movimiento que percibe el observador. Este mapa de vectores se obtiene mediante el algoritmo EMJ (Estimación de Movimiento Jerárquica) que será descrito en el siguiente apartado de forma precisa; se decide utilizar este método de estimación por suponer una alternativa intermedia, en términos de carga computacional, entre la estimación de movimiento por *block-matching* y el cálculo del *flujo óptico*.

El primer clasificador, con un umbral λ_1, determina a partir del módulo del vector correspondiente a cada MB, si éste presenta mucho/poco movimiento, obteniendo así la primera caracterización del MB.

Por otro lado, a partir del mapa de vectores generado por EMJ se estima el movimiento característico de la cámara en la secuencia correspondiente. El objetivo es catalogar los MBs según el grado de interés, y, para conseguirlo, se incluye una etapa de compensación del movimiento de cámara, para relativizar el movimiento percibido con respecto al de cámara y averiguar qué agente causa la sensación de movimiento de la escena, si los objetos o la cámara; se asume que aquellos objetos cuyo movimiento difiera del que realiza la cámara se consideran relevantes o importantes desde el punto de vista subjetivo. El algoritmo encargado de generar el mapa de vectores característico del movimiento de cámara se basa en [26].

Para terminar, las magnitudes de los vectores del mapa compensado se evalúan mediante el clasificador λ_2, generando un mapa binario que delimita el objeto del fondo. De modo que aquellos MBs con vectores de movimiento pequeños o nulos se corresponden con el fondo del plano, mientras aquellos que presentan un grado de movimiento mayor se corresponden con la región de interés. En definitiva, la salida del sistema completo consiste en una combinación de ambos mapas binarios obtenidos, que da lugar a una clasificación cuaternaria que sigue el criterio recogido en la tabla anterior, donde se asigna un nivel de distorsión determinado en cada caso.

Es necesario señalar que el sistema implementado no cubre el criterio completo descrito, sino que se centra en la detección de zonas más susceptibles de aplicar distorsión elevada, es decir, zonas de mucho movimiento y que no son consideradas región de interés. Con este objetivo particular, se describe en los apartados posteriores cada algoritmo implementado y módulo añadido para refinar la solución obtenida.

Capítulo 4

Algoritmo de clasificación de movimiento

4.1 Algoritmo de estimación de movimiento jerárquica

La incorporación de consideraciones perceptuales en codificación de vídeo requiere la delimitación de regiones de interés de la manera más fidedigna posible con respecto a la valoración subjetiva del observador de cada zona del plano. En el sistema propuesto, la base de esa segmentación en zonas de interés es el mapa de vectores calculado por el algoritmo de estimación de movimiento al que se recurre. Al analizar plano a plano el movimiento que percibe un observador, se requiere que la extracción del mapa local de vectores sea lo más parecido al movimiento real presente en la escena. Para seleccionar el algoritmo más adecuado, es necesario conocer las técnicas que se encuentran disponibles.

Los algoritmos tradicionales denominados de *block-matching*, se basan en la búsqueda de un bloque del plano de referencia, dentro de una ventana de búsqueda predefinida, que minimice una función de coste basada en un criterio

determinado, con respecto al bloque original del plano actual. Este tipo de algoritmos consigue eliminar redundancia temporal, y, sólo en ciertos casos consigue representar el movimiento real de la escena, casos como el movimiento de traslación. El criterio que determina la elección del bloque puede ser uno de los siguientes, entre otros, todos ellos relativos al bloque original y al de referencia:

- *Sum of Absolute Differences*, SAD.
- *Sum of Squared Differences*, SSD.
- *Mean Absolute Error*, MAE.
- *Mean Squared Error*, MSE.
- *Mean of Absolute Differences,* MAD.

Por otro lado, los algoritmos de flujo óptico [25] extraen una representación muy acertada y fiel del movimiento real de una escena, pero con un incremento del coste computacional considerable, pues consiste en un campo vectorial donde cada vector representa la velocidad instantánea de cada píxel.

Por tanto, para cubrir los requerimientos de nuestro sistema se recurre a un método alternativo, próximo al flujo óptico, que reduce la carga computacional necesaria, y que también proporciona un mapa de vectores eficiente, la estimación de movimiento jerárquica (EMJ). A continuación se realiza una presentación del algoritmo EMJ utilizado en este sistema, y se incluye una descripción detallada de la configuración del mismo. Para completar este apartado, se muestra la primera versión de la clasificación obtenida a partir del mapa de vectores generado por el algoritmo de estimación seleccionado.

4.1.1 Descripción del algoritmo EMJ

La búsqueda jerárquica realiza la selección del bloque de referencia en varios pasos o niveles jerárquicos. La diferencia entre unos niveles jerárquicos y otros es el tamaño de bloque utilizado en la estimación, pues, el tamaño del bloque (*BS*, *Block Size*) disminuye de forma logarítmica de la capa más baja hacia la superior. Así, en caso de que se trabajase a nivel de MB y se utilizaran tres capas jerárquicas, el tamaño del bloque para j=0 sería de 64x64 píxeles, mientras el correspondiente a la siguiente capa sería de 32x32, teniendo finalmente un tamaño de bloque de 16x16 para el nivel j=2, como se puede observar en la Figura 4.1.

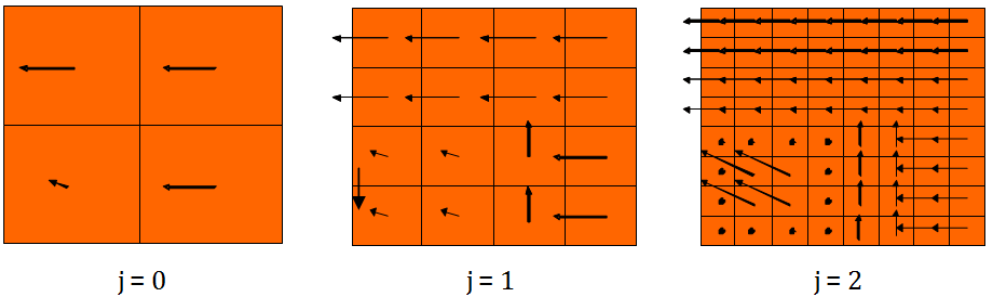


Figura 4.1. Mapa de vectores de movimiento para tres niveles jerárquicos.

Por su parte, el área de búsqueda (*SR*, *Search Range*) es el mismo en todos los niveles, cuya configuración es necesario determinar; además, independientemente de la capa en la que se encuentre el algoritmo, en el área se realiza búsqueda exhaustiva del MB adecuado; en el caso de la primera capa jerárquica, se realiza una estimación burda del movimiento, pues el tamaño del bloque utilizado es grande, y a medida que se aumenta el nivel, la aproximación de la estimación se va refinando.

El criterio de búsqueda utilizado es la MAD. En una primera versión del algoritmo se recurrió a la SAD por ser el criterio más recurrente en trabajos de codificación de vídeo, pero, al analizar los resultados obtenidos, se observó que los valores de SAD obtenidos eran muy elevados y muy dispares entre diferentes bloques de referencia para un mismo bloque original a estimar, como consecuencia de la luminancia característica de cada uno de los bloques de referencia. Sin embargo, utilizando la MAD característica del macrobloque se dispone de un valor medio de la diferencia píxel a píxel entre MBs, siendo éste más orientativo acerca del coste de cada solución posible.

Sobre la MAD se realizan una serie de modificaciones, con el fin de conseguir que el mapa de vectores de la capa jerárquica previa contribuya en la decisión actual. De esta manera, el mapa de un nivel jerárquico influye en la consecución del mapa del nivel superior, de modo que la magnitud de los vectores se regula de una capa a la siguiente. Por lo tanto, la función de coste utilizada para un nivel jerárquico "j" se corresponde con la siguiente expresión:

$$Coste(vm^j) = \begin{cases} MAD(vm^j) + \blacksquare \cdot |vm^j|, j = 0 \\ MAD(vm^j) + \alpha_1 \cdot |vm^j - vm^{j-1}| + \alpha_2 \cdot |vm^j|, j > 0 \end{cases} \quad (18)$$

Donde $vm = (vm_x, vm_y)$ representa una solución potencial para el vector de movimiento final, y α_1 y α_2 son parámetros de regularización que es necesario configurar. El término asociado a α_1 compara el vector de movimiento de la capa

jerárquica actual con las estimaciones previas llevadas a cabo en el nivel anterior j-1, penalizando variaciones grandes. De este modo el vector refinado no debe diferir notablemente del obtenido en la capa jerárquica inferior (véase la Figura 4.2).

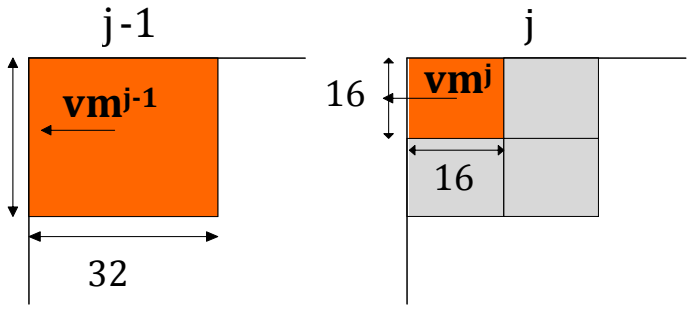


Figura 4.2. Un vector de movimiento vm^j calculado en la capa jerárquica *j* debe asemejarse a vm^{j-1}, el obtenido en la capa jerárquica inferior.

Por su parte, el parámetro α_2 pondera la magnitud del vector de movimiento asignado a la capa actual. Se encarga de penalizar movimientos grandes y así evitar resultados anómalos. Así, este parámetro es de especial importancia en zonas muy homogéneas, donde el valor de la MAD es tan bajo que podría dar lugar a una solución aleatoriamente alejada del (0,0), dando lugar a un vector que no representa el movimiento real de la escena. Y, gracias a α_2, aquellas posibles soluciones de magnitud destacable en regiones conflictivas son descartadas en la elección del vector de movimiento final asociado al MB correspondiente.

Es necesario señalar que la función de coste correspondiente a la primera capa jerárquica tiene asignado un peso independiente en el módulo del vector, como se puede observar en (18); esto se debe a la importancia que tiene el mapa de vectores correspondiente al primer nivel. Un mapa de vectores erróneo en la primera capa tiene como consecuencia un mapa de vectores final no fiable, pues el peso de las capas anteriores en la función de coste hace que el error se arrastre hacia las capas superiores. Por esta razón, se determinó independizar el ajuste del peso de la función de coste correspondiente a la capa inferior, con respecto al coste asociado al resto de capas.

También, hay que destacar que el algoritmo EMJ es compatible tanto con patrón de GOP tipo IP, como con patrones que incluyen planos tipo B. En el caso de que el patrón sea IP, no hay ningún problema con respecto a los planos de referencia en la estimación puesto que, como se señaló en el apartado 2.2.1, el orden de codificación se corresponde con el de presentación, de modo que siempre

se dispone de los planos anteriores para poder aplicar el algoritmo EMJ. Por otro lado, cuando el patrón contiene planos tipo B, la codificación previa de los planos tipo P a los B, da lugar a duda con respecto a la disposición de los planos anteriores necesarios en la estimación. Sin embargo, como los planos B han tenido que ser leídos y almacenados en memoria previamente, se dispone en todo momento de los planos inmediatamente anteriores. Además, lo más coherente parece ser tomar como imagen de referencia la más cercana en el tiempo a la actual, puesto que de esta manera la estimación del movimiento es mucho más fiable. Por lo tanto, el algoritmo EMJ es independiente del patrón de codificación, pues se rige por el orden de presentación de los planos.

4.1.2 Métodos de reducción del coste computacional

Uno de los factores más críticos del algoritmo EMJ es el elevado coste computacional que supone. Aunque no conlleva el tiempo de cómputo asociado al flujo óptico, esta técnica obtiene mapas de vectores de movimiento de cada nivel jerárquico correspondiente a cada plano de la secuencia a analizar. Por tanto, el número de operaciones realizadas por plano es elevado.

En este apartado se proponen un conjunto de medidas que pretenden agilizar el proceso, reduciendo los cálculos por medio de diezmados en lugares concretos del algoritmo o asignando valores predeterminados a ciertos parámetros de configuración que implican una reducción del número global de operaciones necesarias. A continuación, se listan los métodos utilizados.

- Limitar el número de niveles jerárquicos y el tamaño del área de búsqueda

  A parte de los coeficientes α_1 y α_2 de la función de coste (ver expresión (18)), el número de niveles jerárquicos (J) y el tamaño del área de búsqueda (SR) son los parámetros del algoritmo que uno debe fijar a priori. Cabe esperar que sus valores no deban ser elevados dada la cercanía temporal entre las imágenes original y de referencia (una y cuatro unidades de tiempo para representar los patrones IP e IP3B respectivamente), y por tanto, las similitudes entre ellas.

  Estos dos parámetros conllevan una carga computacional importante, pues, en primer lugar, el tamaño del área de búsqueda implica el número

de macrobloques de referencia posibles que debe analizar; por su parte, el número de niveles jerárquicos indica la precisión del algoritmo, por tanto, a mayor número de niveles, mayor cantidad de operaciones.

Al ser configurables, será necesario realizar una serie de pruebas con diferentes configuraciones, para observar las diferencias en el número de operaciones realizadas, así como la calidad de los resultados obtenidos con cada una de ellas, con el fin de establecer un valor predeterminado adecuado en cada caso.

- Limitar el número de puntos de evaluación en la función MAD

  La función MAD empleada como criterio de evaluación en la función de coste puede resultar muy costoso computacionalmente, sobre todo en niveles jerárquicos bajos, pues el tamaño de bloque es grande.

  Por tanto, si en lugar de evaluar la diferencia absoluta entre cada uno de los puntos que alberga el bloque de cada nivel jerárquico, se limita el número de puntos a evaluar mediante un diezmado regular, el número de cálculos realizados se vería reducido de manera considerable. Sin embargo, sería aconsejable realizar el cálculo completo de la MAD en la capa jerárquica superior, cuyo tamaño de bloque correspondiente es el menor, pues ya implica un número de píxeles lo suficientemente bajo.

  Este método propuesto de reducción hay que evaluarlo con las pruebas correspondientes, puesto que la calidad de la solución obtenida puede verse afectada considerablemente, al fin y al cabo se está prescindiendo de información a la hora de evaluar la validez de un bloque de referencia; por ello, es necesario encontrar una solución de compromiso, de modo que la información excluida no sea de relevancia y no implique pérdidas de calidad.

- Limitar el número de bloques de referencia a evaluar dentro del área de búsqueda

  Además de configurar el tamaño del área de búsqueda de forma adecuada, se puede tomar sólo un subconjunto de bloques que ocupan dicho área para evaluar, reduciendo el número de comprobaciones. Esto supondría un diezmado similar al presentado en la medida anterior, más

apropiado en niveles jerárquicos inferiores, donde se pretende obtener groso modo una estimación del movimiento real de la escena.

Para finalizar, señalar que si se aplican las propuestas anteriores de forma simultánea, se puede reducir aún más el coste computacional. Pero la combinación de estas medidas debe realizarse con sutileza, siempre buscando un equilibrio entre los cálculos que supone y la pérdida de calidad que pueda implicar.

4.1.3 Configuración del algoritmo

El algoritmo de estimación de movimiento cuenta con una serie de parámetros configurables, cuyo valor es necesario predeterminar de forma eficaz para conseguir un compromiso entre el coste computacional y la calidad del mapa de vectores de movimiento obtenido.

Se tratan de los parámetros de regularización presentes en la función de coste del algoritmo, α_1 y α_2, el número de capas jerárquicas (J) y el tamaño del área de búsqueda o *SR*. Los pesos correspondientes a la función de coste tienen una influencia muy importante sobre el mapa de vectores generado, de modo que la asignación de valor a los mismos es más relevante frente a la configuración del número de niveles jerárquicos y el área de búsqueda. Por esta razón, para poder lanzar pruebas para configurar los parámetros prioritarios, se determina fijar los valores de J a [redacted] niveles y SR a un tamaño de [redacted] píxeles, considerando ésta una configuración suficientemente robusta y equilibrada con respecto al coste computacional y la calidad del mapa resultantes. A continuación se describen las pruebas realizadas, y se incluyen algunas capturas como muestra de resultados.

(1). Prueba primera capa jerárquica

En primer lugar, como ya se explicó en el apartado de descripción del algoritmo, la primera prueba realizada para ajustar el valor de un parámetro se corresponde con el peso asignado en la función de coste de la primera capa jerárquica, que tiene un valor de [redacted] y al que denominamos α_0. Se lanzó una prueba en la que el algoritmo sólo calculaba el mapa correspondiente a la capa jerárquica inferior cuyo tamaño de MB es de [redacted] píxeles, utilizando una SR de [redacted]. La prueba estaba compuesta por dos grupos de secuencias de vídeos, un grupo compuesto por secuencias de tamaño CIF (352x288) y otro compuesto por secuencias de

tamaño SD (720×576 ó 704x480); algunas de las secuencias originales utilizadas en estas pruebas se adjuntan en el DVD (véase Anexo I), sólo se han incluido las más representativas. Por su parte, el rango de pesos evaluado fue [redacted]. También, de forma adicional, se añade el valor 0 en alguna secuencia para observar la importancia del término que se descartaba. Las conclusiones extraídas al respecto son las siguientes:

- La importancia de este término de la función de coste se refleja en la estimación de MBs con características de textura homogénea de manera eficaz. Cuando el peso del vector de movimiento en la función de coste es bajo o nulo, en zonas uniformes aparecen vectores de módulo elevado que no suponen una representación fiel del movimiento real de la escena. En las siguientes imágenes, se puede observar cómo la presencia de los vectores en la función de coste, por mínimo que sea su coste ([redacted] en este caso), consigue mejorar el mapa de vectores correspondiente, eliminando un porcentaje considerable de vectores erróneos.

Figura 4.3. Mapa de vectores para α_0=[redacted].
Secuencia "Ice Age" (720x576).

Figura 4.4. Mapa de vectores para α_0=[redacted].
Secuencia "Ice Age" (720x576).

- Por otro lado, se busca conseguir un mapa de movimiento lo más uniforme posible, de modo que se distingan regiones con movimiento elevado y se pueda apreciar levemente una segmentación característica del grado de movimiento presente en la escena. Para conseguirlo, hay que ajustar α_0, una vez vista la necesidad de utilizarlo. Así, tras analizar los mapas resultantes para cada valor de α_0, se concluye que [redacted] es el valor más apropiado, pues, como se pueden observar en las siguientes capturas, donde la espada del samurái se desplaza hacia abajo, el movimiento queda mejor definido.

Figura 4.5. Mapa de vectores para α_0=[redacted].
Secuencia "Último Samurai" (720x576).

Figura 4.6. Mapa de vectores para α_0=[redacted].
Secuencia "Último Samurai" (720x576).

(2). Prueba selección parámetros α1 y α2

Una vez asegurado el cálculo eficaz del movimiento en la capa inferior del algoritmo, se deben ajustar el resto de parámetros de la función de coste, de modo que a partir de la primera aproximación del movimiento, las siguientes capas refinen el mapa de forma adecuada. Como ya se destacó anteriormente, los pesos correspondientes a la función de coste son de gran importancia, por ello se realiza una prueba exhaustiva de selección de valores para α_1 yα_2 con la que poder demostrar la viabilidad del sistema, sin llevar a cabo una batería de pruebas más densa con otros valores de J, de SR y con los métodos propuestos para reducir el coste computacional.

Con objeto de encontrar la combinación óptima de α_1 y α_2, se establece la siguiente configuración del algoritmo:

- Niveles Jerárquicos (J): [redacted]
- Área de búsqueda (SR): [redacted]
- Diezmado en SR: [redacted]
- Diezmado en cálculo MAD: [redacted]

Las opciones de diezmado regular existen en el algoritmo, a pesar de que para esta batería de pruebas no se han utilizado, pues la configuración debía ser la más robusta. Las otras dos magnitudes se fijan a los valores determinados anteriormente.

Esta prueba se compone de dos categorías de vídeos, una de vídeos con fondo estático (véase "football.cif" a modo de ejemplo), donde la cámara no se mueve pero el objeto de interés sí, y otra de vídeos con fondo en movimiento (véase "bohemia.cst"), en los que la cámara sigue al objeto de interés, de modo que el objeto de interés parece inmóvil con respecto a la cámara y el movimiento se concentra en el fondo. En función del conjunto a tratar, el mapa obtenido se interpreta de una manera u otra; así, en el caso de fondo estático, la zona de interés es aquella cuyos vectores sean no nulos, mientras que en el caso de fondo en movimiento, interesa la región con movimiento nulo, pues se tratará del objeto al que sigue la cámara. Señalar que a la categoría de fondo estático pertenecen 8 secuencias, mientras, a la segunda categoría pertenecen 6 secuencias de vídeos, y, en ambos casos, la mitad de secuencias tienen tamaño CIF y la mitad restante SD.

Los parámetros α_1 y α_2 varían dentro de los siguientes rangos de valores para determinar el mapa resultante para cada una de las combinaciones posibles:

- α_1=[redacted]
- α_2=[redacted]

Para cada resultado obtenido se ha realizado un análisis de calidad subjetiva, determinando así el par de valores que ofrece los mejores resultados para cada uno de los vídeos, comprobando que en el mapa quedan bien definidas las figuras en movimiento o estáticas, según corresponda, así como que no aparezcan vectores incoherentes en determinadas regiones.

Tras realizar esta prueba se alcanzan las siguientes conclusiones:

- En regiones de la imagen con mucho detalle el algoritmo funciona correctamente. En este caso, los términos asociados a α_1 y α_2 no son determinantes en la elección del mejor bloque de referencia, sino que todo el peso de la decisión lo lleva el valor de MAD correspondiente.

- Los parámetros α_1 y α_2 influyen en caso de valores de MAD bajos, pues se trata de zonas con un grado de detalle insuficiente, donde el vector de la capa anterior y el vector asociado al mismo cobran mayor importancia. Es en este tipo de regiones aparecen dos casos diferenciados:

    - En el interior de objetos grandes en movimiento que tienen textura uniforme, en función de los valores que tomen estos dos parámetros, los vectores de movimiento valen mayoritariamente cero y/o surgen vectores incoherentes con el movimiento real del objeto en cuestión (*outliers*). Por tanto, surge un problema de rellenado de huecos o de *outliers*.
    - Cuando el fondo es uniforme y existe movimiento de cámara, el algoritmo no puede detectar el movimiento, debido a las restricciones impuestas en el mismo.

    Como solución a estas dos situaciones se determina priorizar el vector de movimiento (0,0) a través del término α_2, considerándose un compromiso razonable en aquellas regiones en las que el algoritmo es incapaz de estimar el movimiento real. Por tanto, aquí queda reflejada la necesidad de algún tipo de procesado que mejore el mapa de vectores generado, rellenando aquellos objetos cuyo movimiento se refleja en sus bordes puesto que su interior está hueco.

- Las observaciones realizadas con respecto a la relación de tamaño entre los valores se describen en las conclusiones restantes. Así, en primer lugar, cuando se cumple que $\alpha_1 \geq \alpha_2$, los vectores de movimiento del interior de áreas uniformes no son nulos, de modo que desaparece el problema de los huecos; pero esta mejora hace que la delimitación del objeto con respecto al fondo no sea completamente fidedigna, pues, como se puede observar en la Figura 4.7, según el mapa de vectores, la silueta del personaje se sitúa más hacia la izquierda en el plano de lo que en realidad está.

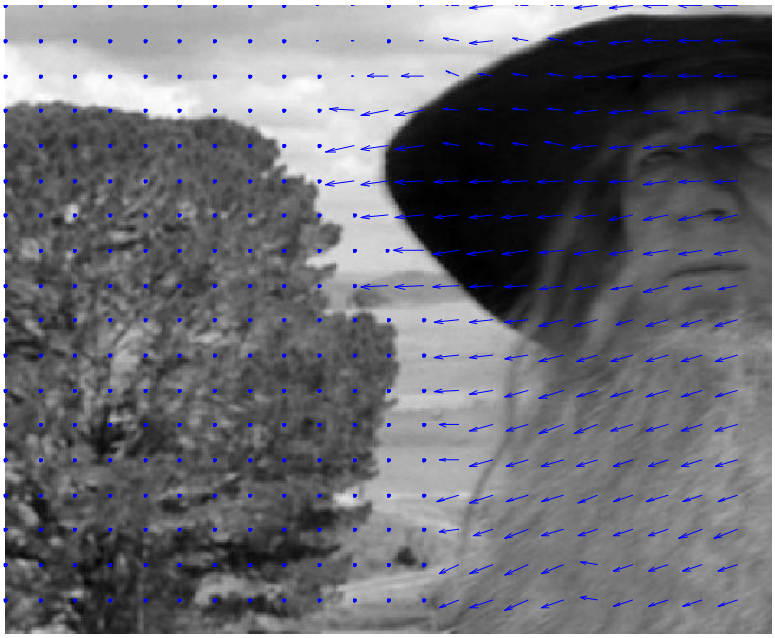


Figura 4.7. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted]. Secuencia "LOTR" (352x288).

- Por otro lado, cuando $\alpha_1 < \alpha_2$, sucede el caso contrario a lo comentado antes: el algoritmo segmenta mejor, pero aparecen huecos en el interior de objetos de textura uniforme. En la Figura 4.8 se puede observar la presencia de vectores nulos en el interior del gorro o de la barba, pero el objeto queda bien delimitado por el mapa de vectores.

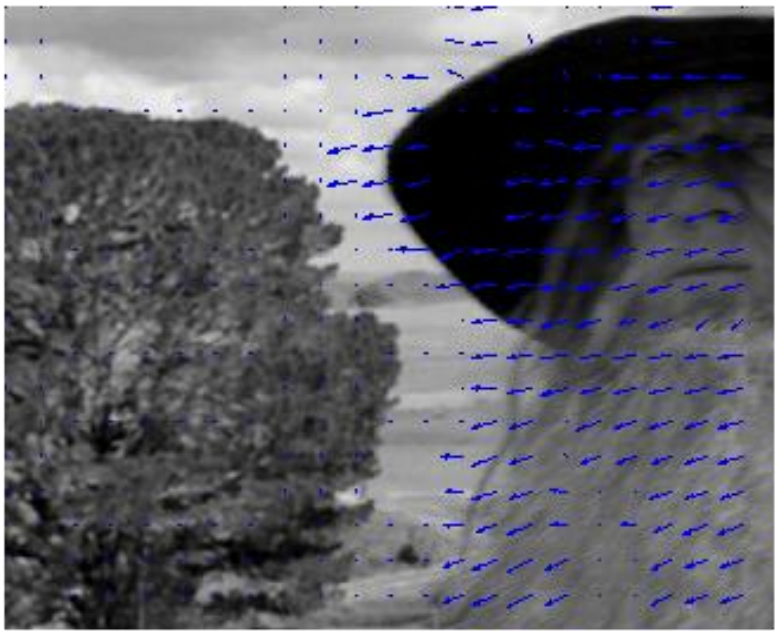


Figura 4.8. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted]. Secuencia "LOTR" (352x288).

- El término α_2 debe ser suficientemente grande para evitar que en regiones uniformes aparezca gran variedad de vectores que apunten a zonas distintas; pero su valor tampoco debe ser excesivo puesto que podría agravar el problema de la aparición de huecos en este tipo de regiones.

En definitiva, en base a las conclusiones extraídas, se considera adecuado determinar este rango de valores de α_1 y α_2: [redacted] ≤ α_1 ≤ [redacted], [redacted] ≤ α_2 ≤ [redacted]. Los resultados obtenidos relativos a las posibles combinaciones de valores incluidos en este rango son generalmente correctos, aunque no todos los pares de valores son eficientes en todas las secuencias utilizadas. Por ello, y por las observaciones descritas anteriormente en las conclusiones, se considera más acertado el caso en el que $\alpha_1 < \alpha_2$, pues es más importante conseguir una buena capacidad de segmentación que rellenar mejor los huecos dentro de zonas uniformes. La elección se debe a que el objetivo del algoritmo es detectar aquellas zonas con un movimiento elevado para codificarlas de manera más burda que el resto.

Por lo tanto, de los casos que cumplen la condición elegida ($\alpha_1 < \alpha_2$) se proponen dos pares de valores para completar la función de coste:

- α1 = [redacted] y α2 = [redacted]
- α1 = [redacted] y α2 = [redacted]

Los mapas de vectores obtenidos con cada par de valores son muy parecidos. Sin embargo, difieren en lo siguiente: en el caso de la primera combinación de valores, se consiguen rellenar mejor los huecos de regiones homogéneas en movimiento, aunque aumenta la probabilidad de que aparezcan *outliers*, debido a que el valor de α_2 no es lo suficientemente grande en algunas ocasiones en los que no consigue hacer prevalecer el vector nulo. Todo esto se puede observar en la Figura 4.9, donde se aprecian *outliers* en el cielo, mientras la región homogénea señalada en el suelo no tiene huecos, dando lugar a una asignación de vectores más uniforme que la del plano en la Figura 4.10, donde, aunque los *outliers* son eliminados, los huecos en áreas de textura homogénea son más abundantes.

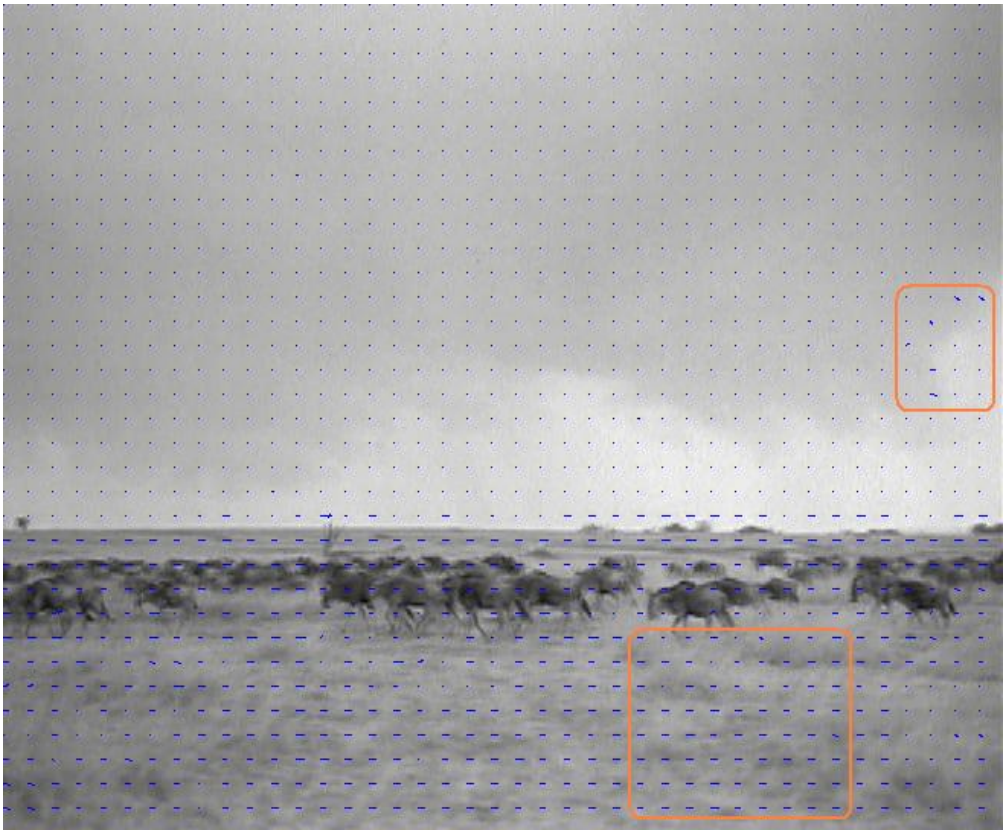


Figura 4.9. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted].
Secuencia "África" (656x544).

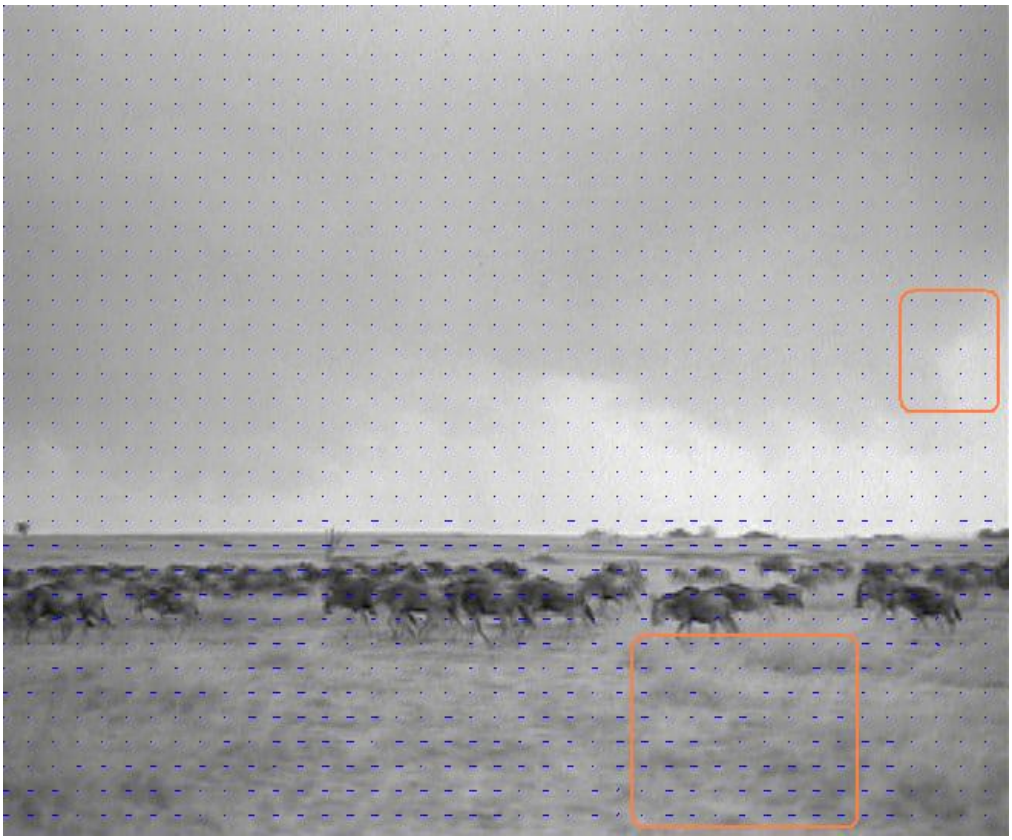


Figura 4.10. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted].
Secuencia "África" (704x480).

Por otro lado, para el segundo caso, se obtiene una mejora en la consistencia temporal del mapa de vectores, es decir, las diferencias en los vectores de un plano al siguiente son menores, suavizando las variaciones entre planos consecutivos, consiguiendo además un mayor grado de uniformidad en los vectores asociados a regiones concretas del plano. Como se puede apreciar en las siguientes imágenes con el par de valores α_1 = [redacted] y α_2 = [redacted], los vectores se agrupan de manera más uniforme, aunque para ambos mapas el movimiento de giro presente en esta secuencia queda reflejado.

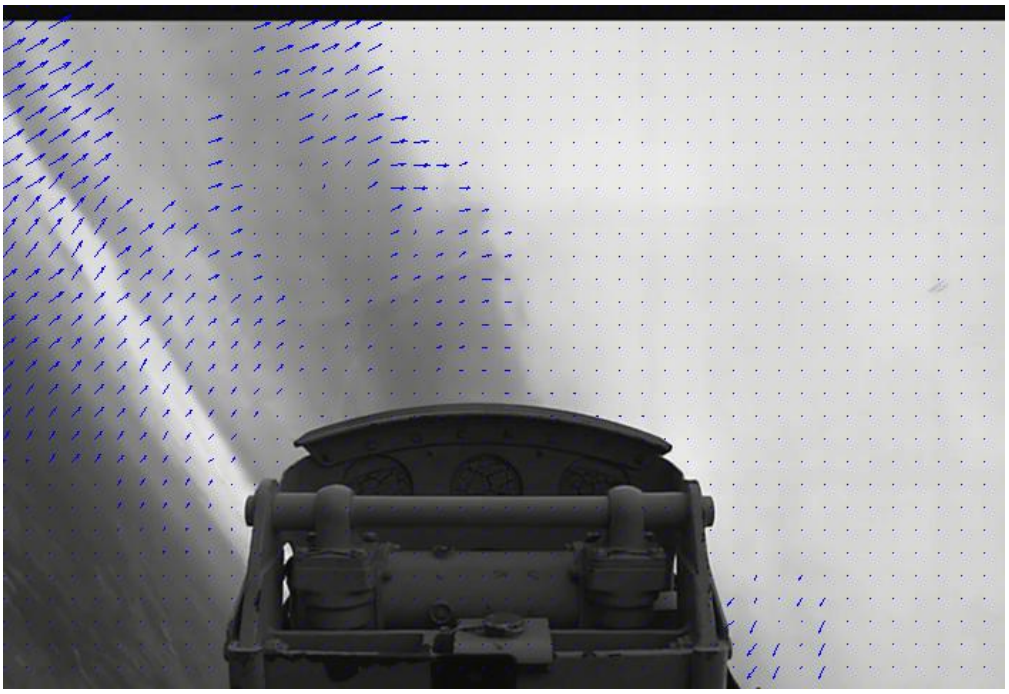


Figura 4.11. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted].
Secuencia "Airshow1" (704x480).

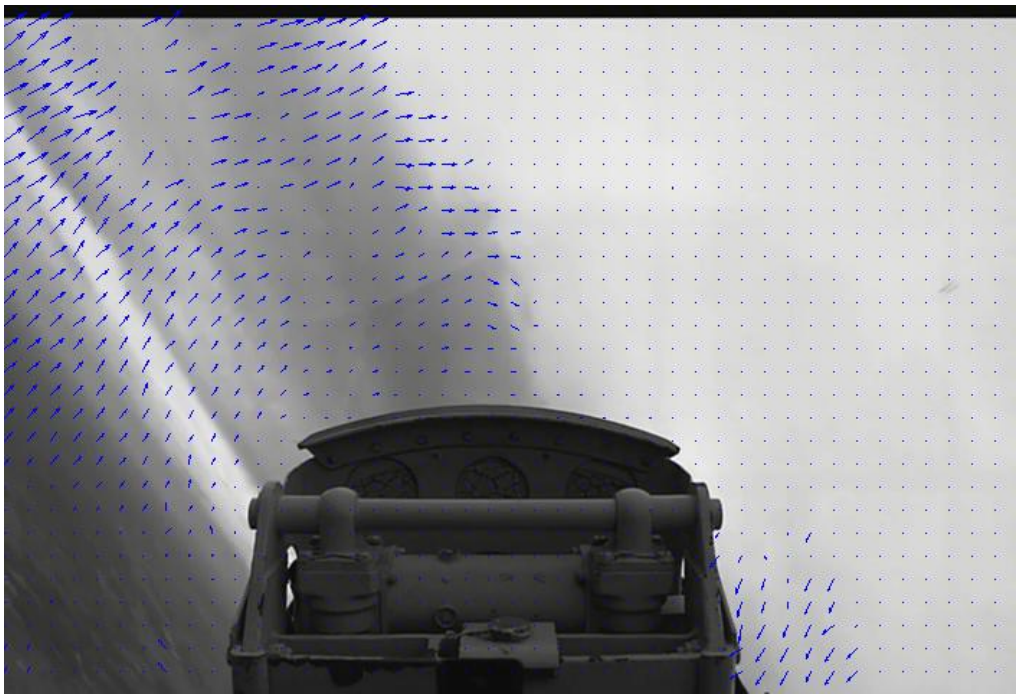


Figura 4.12. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted].
Secuencia "Airshow1" (704x480).

(3). Prueba cuarta capa jerárquica

A la vista de los resultados obtenidos con la configuración robusta prefijada, se determina realizar una prueba para corroborar la eficiencia de la selección de [redacted] capas jerárquicas en las pruebas anteriores y poder concluir la etapa de configuración robusta del algoritmo.

Se aumenta el número de niveles jerárquicos utilizados a 4, mientras el resto de parámetros se mantienen a sus valores por defecto, y los pesos de la función de coste se fijan a α_1 = [redacted] y α_2 = [redacted]. Las secuencias a analizar se corresponden con un subconjunto de la batería de pruebas anterior: 3 con fondo en movimiento y 3 con fondo estático.

Los resultados obtenidos indican que [redacted] capas jerárquicas son suficientemente robustas para realizar una ejecución pesada, pues no existen apenas diferencias en los mapas de vectores obtenidos al configurar el algoritmo con [redacted] niveles jerárquicos. Por lo tanto, utilizando [redacted] capas el algoritmo funciona eficientemente y además supone un coste computacional menor que recurrir a un número de capas superior. Las siguientes figuras se consideran una muestra de las similitudes entre los mapas obtenidos en cada caso.

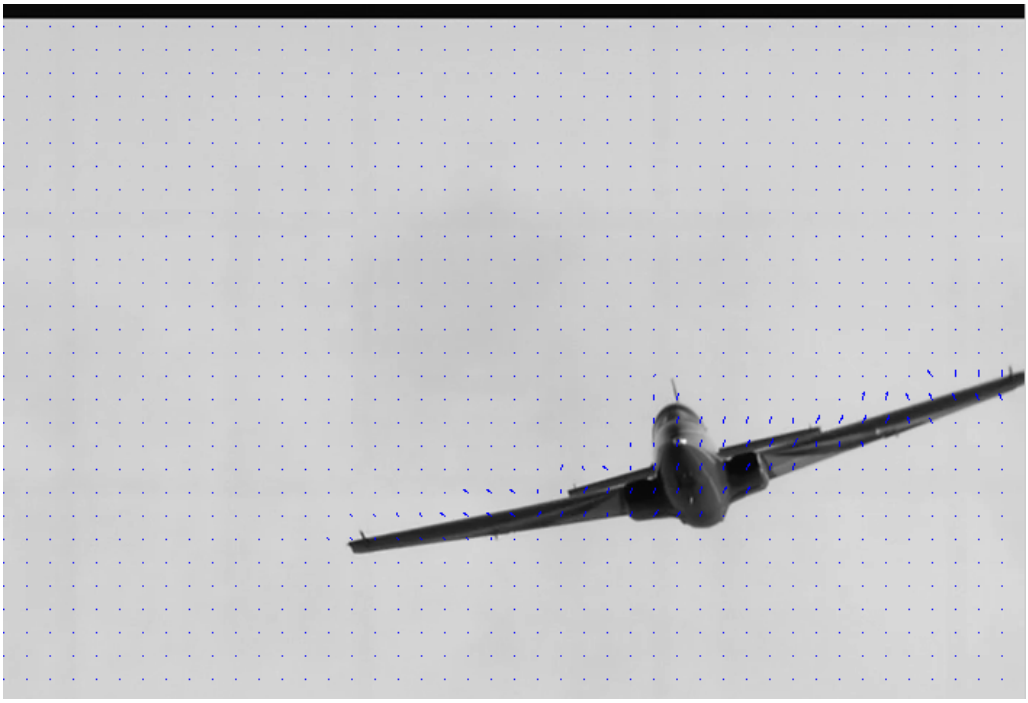


Figura 4.13. Campo de vectores de movimiento. J=[redacted], α_1 = [redacted], α_2 = [redacted]. Secuencia "Airshow2" (704x480).

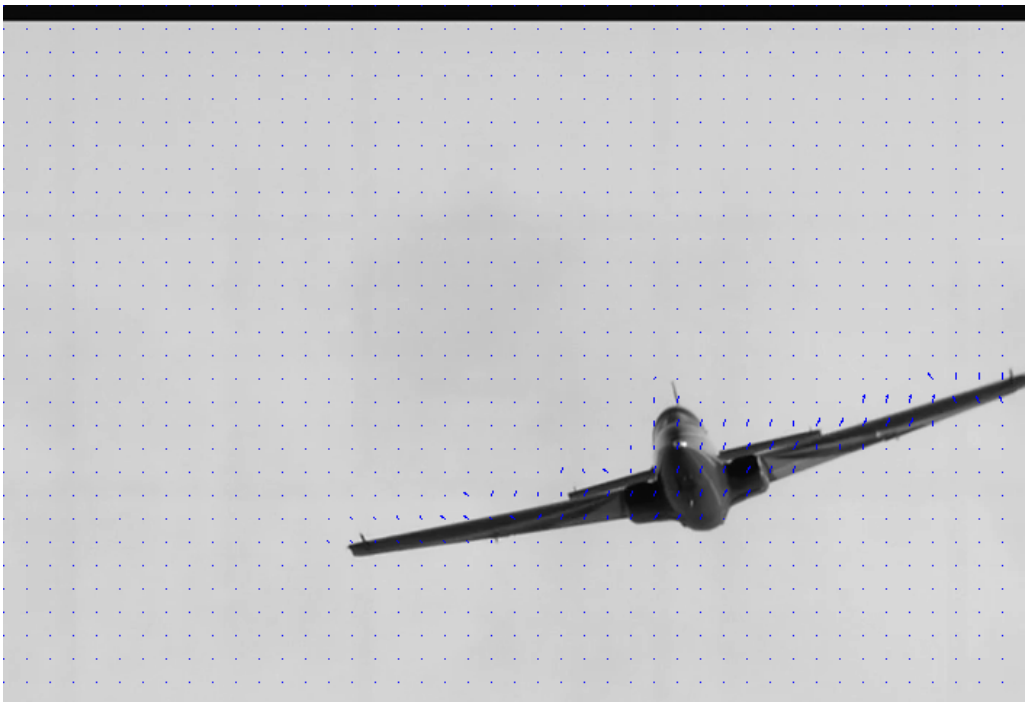


Figura 4.14. Campo de vectores de movimiento. J=[redacted], α_1 = [redacted], α_2 = [redacted].
Secuencia "Airshow2" (704x480).

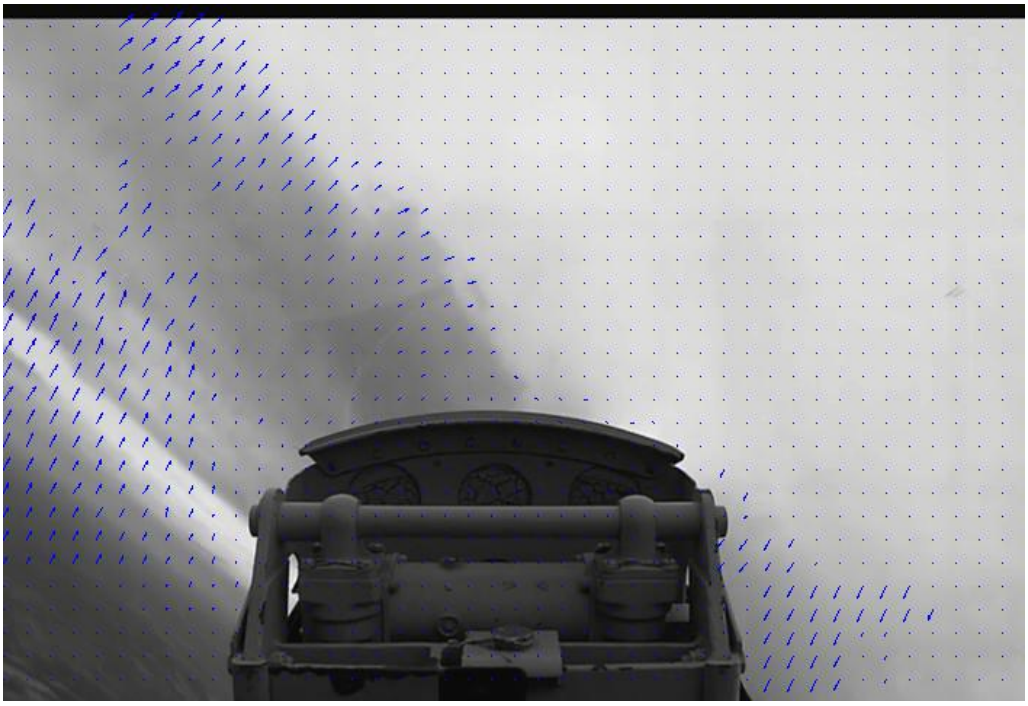


Figura 4.15. Campo de vectores de movimiento. J=[redacted], α_1 = [redacted], α_2 = [redacted].
Secuencia "Airshow1" (704x480).

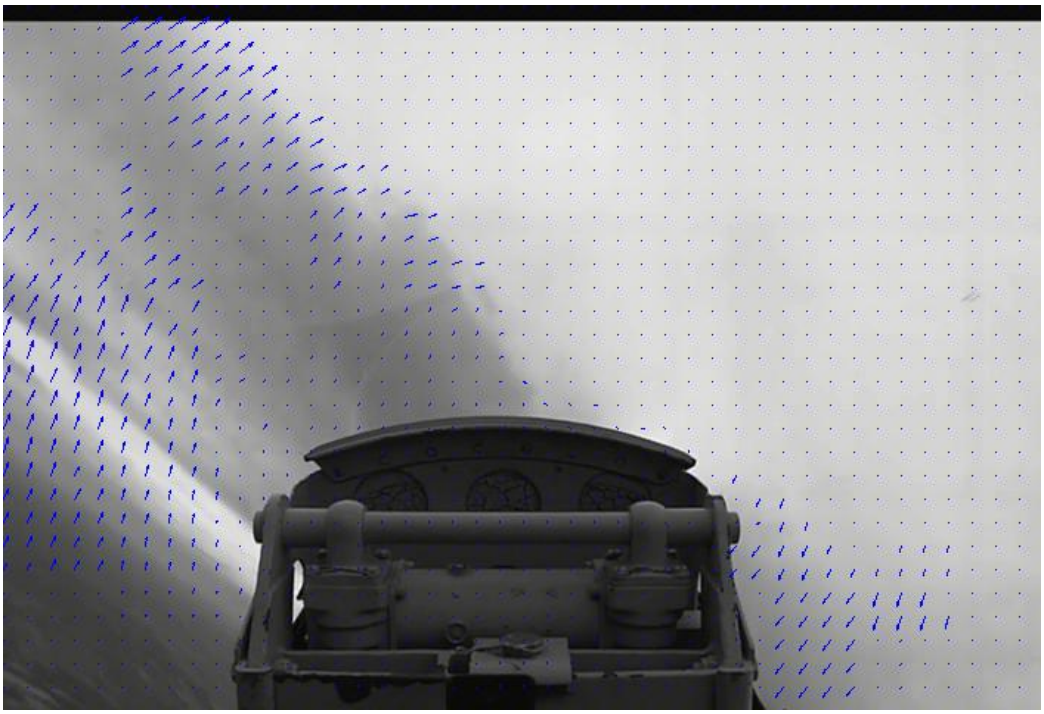


Figura 4.16. Campo de vectores de movimiento. J=[redacted], α_1 = [redacted], α_2 = [redacted].
Secuencia "Airshow1" (704x480).

4.2 Clasificador binario de movimiento λ_1

Una vez determinada la configuración del algoritmo EMJ, es necesario continuar con la implementación del sistema de enmascaramiento por movimiento. El siguiente paso consiste en comprobar la viabilidad del mismo, realizando una primera aproximación de clasificación del movimiento, y evaluando la decisión binaria resultante.

Mediante el diseño de un clasificador sencillo de carácter binario, se indica si un MB tiene mucho movimiento (1) o poco (0), según el módulo de su vector. El umbral que asigna la etiqueta de mucho/poco movimiento a cada MB se denomina λ_1 y para encontrar su valor más adecuado hay que lanzar una batería de pruebas compuesta por secuencias con características de movimiento diferentes, variedad de tamaños de vídeo, y un rango amplio de valores de λ_1.

En concreto, la batería de pruebas está compuesta por el grupo de secuencias de tamaños CIF y SD utilizadas en la configuración de los parámetros del algoritmo EMJ. La novedad de esta prueba consiste en la introducción de las mismas secuencias de vídeo escaladas a tamaño QCIF (176x144), con el fin de

observar la influencia del tamaño en la clasificación, y determinar un umbral adecuado al mismo. Adicionalmente, se comprueba la eficiencia del algoritmo EMJ para tamaños de vídeo tan pequeños, en los que un MB de 16x16 supone una región amplia de la escena.

La configuración del algoritmo utilizada para llevar a cabo las pruebas es la siguiente:

- Niveles Jerárquicos (J): [redacted]
- Área de búsqueda (SR): [redacted]
- Diezmado en SR: [redacted]
- Diezmado en cálculo MAD: [redacted]
- Parámetro de regularización α_1: [redacted]
- Parámetro de regularización α_2: [redacted]

Además el rango de valores de prueba de λ_1 se recoge en la siguiente tabla para cada una de las resoluciones de las secuencias disponibles. Así, podemos observar que los umbrales evaluados en el caso de secuencias CIF son la mitad de los asignados a vídeos SD, mientras la magnitud de los umbrales para QCIF es la cuarta parte.

Formato Vídeo	**SD**	**CIF**	**QCIF**
Umbral λ_1	[redacted]	[redacted]	[redacted]

Tabla 4.1. Umbrales λ_1 de prueba

En general, la clasificación en todo el conjunto de secuencias es satisfactoria, aunque los umbrales de mayor magnitud restringen mucho la clasificación (véase Figura 4.17) en cada caso; también, en el caso de los umbrales más bajos, las regiones con un grado de movimiento destacado son más abundantes, de modo que el mapa generado no va acorde con la clasificación objetivo que se busca, pues no permiten delimitar las regiones de interés de la escena (véase Figura 4.18).

Las capturas que se adjuntan a continuación y en el resto de capítulos de este proyecto se corresponden con planos de las secuencias de prueba sobre los que se dibuja el mapa de vectores característico junto con la clasificación, representada por los MBs coloreados en naranja que se corresponden con regiones del plano catalogadas como con mucho movimiento.

Figura 4.17. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted]+ clasificador λ_1 = [redacted].
Secuencia "Bohemia" (704x576).

Figura 4.18. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted]+ clasificador λ_1 = [redacted].
Secuencia "Bohemia" (176x144).

A continuación se muestran una serie de ejemplos de clasificación obtenidos. Las cuatro primeras figuras se corresponden con planos de secuencias en las que se realiza un seguimiento a un objeto de interés, de modo que el movimiento se sitúa en el fondo; el resto de imágenes pertenecen a vídeos en los que no hay movimiento de cámara, sino que es el objeto de interés el que presenta un movimiento acentuado.

Con respecto al valor óptimo del umbral para cada tipo de vídeo, es necesario señalar que aquellos de tamaño SD presentan mejores resultados cuando λ_1 = [redacted] y λ_1 = [redacted]. Por su parte, los mapas binarios de las secuencias CIF han

resultado razonablemente buenos ante umbrales de [redacted] y [redacted] píxeles por lo general. Finalmente, los mapas de los vídeos QCIF más acertados se corresponden con un umbral de [redacted] píxeles. Las capturas que se encuentran a continuación no son sólo representativas de los mejores resultados obtenidos, sino también de los defectos que presenta el clasificador en casos concretos y que se comentan con más detalle posteriormente.

Figura 4.19. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted]+ clasificador λ_1 = [redacted]. Secuencia "Bohemia" (704x576).

Figura 4.20. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Bohemia" (176x144).

Figura 4.21. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Bus" (352x288).

Figura 4.22. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Bus" (176x144).

Figura 4.23. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Football" (352x288).

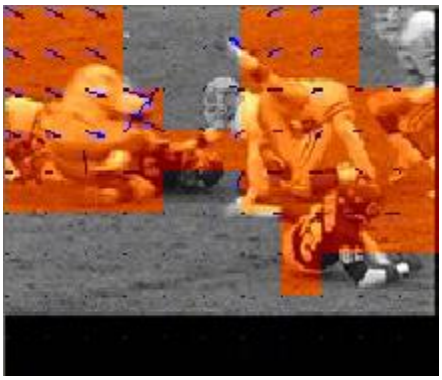


Figura 4.24. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Football" (176x144).

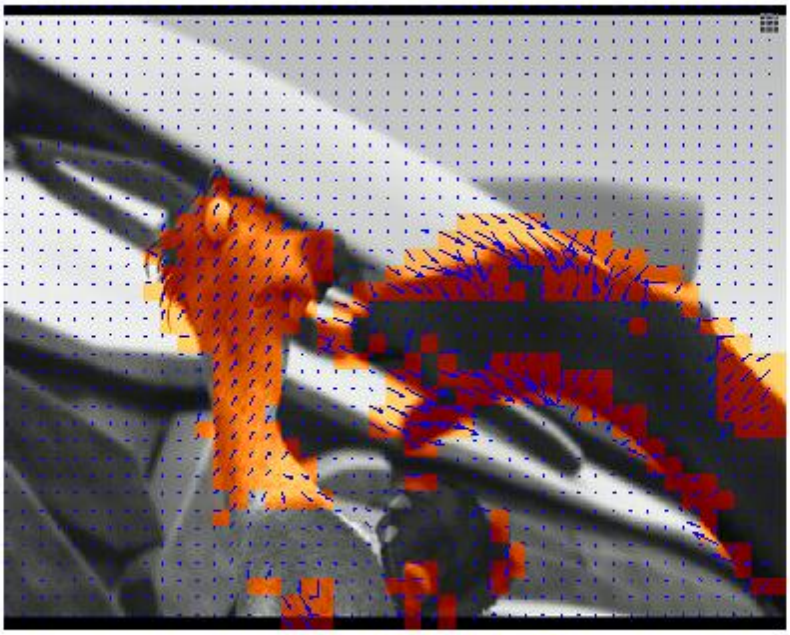


Figura 4.25. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Ice Age" (720x576).

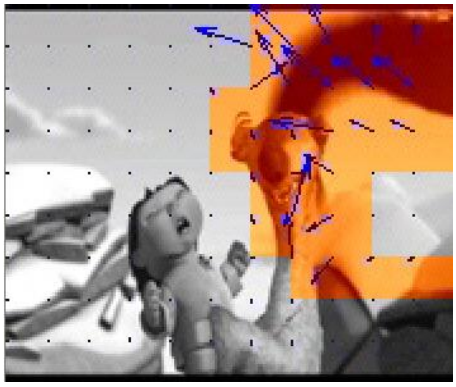


Figura 4.26. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Ice Age" (176x144).

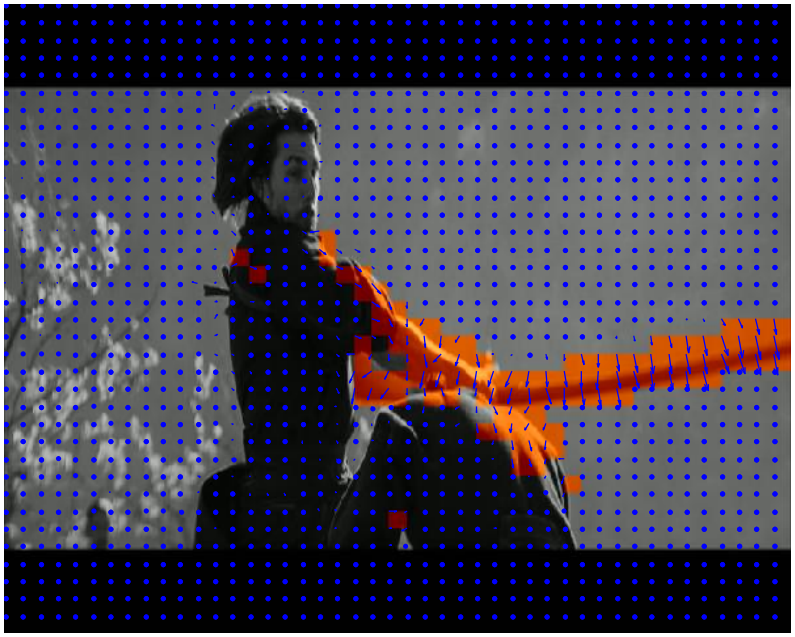


Figura 4.27. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Último Samurai" (720x576).

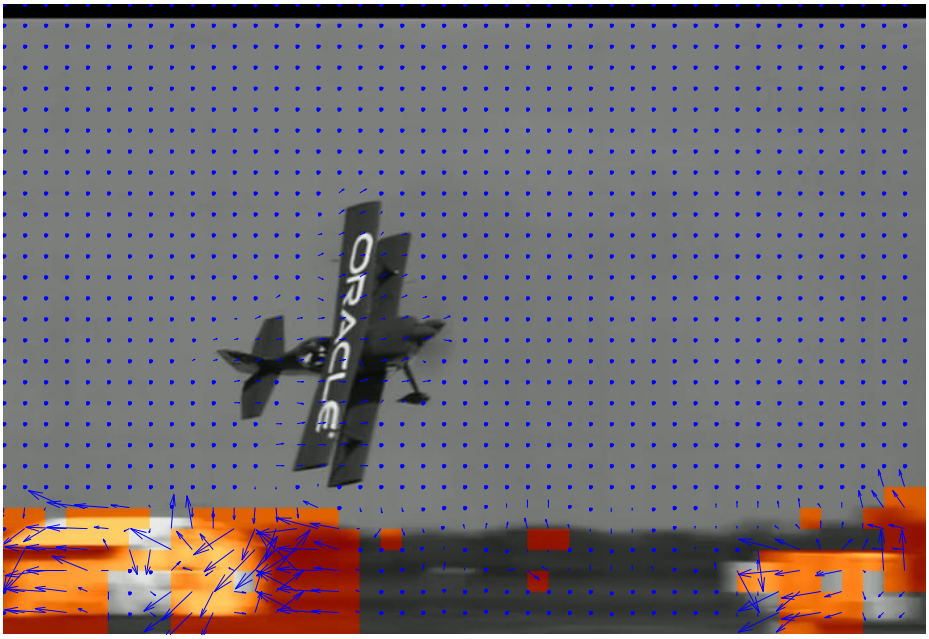


Figura 4.28. Campo de vectores de movimiento. α_1 = [redacted], α_2 = [redacted] + clasificador λ_1 = [redacted]. Secuencia "Airshow3" (704x480).

En general, se puede observar una clasificación eficiente en todos los casos, pues las regiones con mucho movimiento son detectadas. Sin embargo, existen áreas que deberían ser consideradas con mucho movimiento y no lo son, o viceversa; esto se produce en las siguientes situaciones:

- En primer lugar, en las regiones que presentan características de textura homogénea no se detecta el movimiento de la cámara, pues los vectores asociados a dichos MBs son nulos. Esto se aprecia en las paredes de la secuencia "*Bohemia*" o en el cielo de "*Airshow3*" (véase "bohemia_morphoOFF_Thx.avi" disponible en el DVD adjunto). Se asigna el vector (0,0) porque se considera el mejor del peor de los casos posibles, es decir, es más conveniente considerar el vector (0,0) que la posibilidad de obtener un vector de módulo considerable que apunte a un MB de referencia sin sentido alguno.
- Por otro lado, en la Figura 4.25 se puede observar cómo la trompa del mamut sólo presenta movimiento en los bordes de ésta. Este caso se produce en objetos grandes en movimiento cuyo interior es uniforme; la asignación del vector se rige por la política del caso anterior y sólo MBs con información de bordes pueden ser clasificados correctamente.
- El siguiente caso anómalo consiste en la aparición de MBs aislados clasificados a 1, que deterioran la consistencia espacial de la clasificación. Como muestra de ello, se dispone de la Figura 4.21 y 4.27. La influencia de estos bloques aislados en la calidad subjetiva de la secuencia codificada no es del todo predecible, aunque parece lógico considerar la eliminación de los mismos para conseguir un mapa binario más uniforme. Si se observa la secuencia "bus_morphoOFF_Thx.avi" disponible en el DVD adjunto, se puede percibir la presencia de MBs aleatorios del interior del autobús clasificados de manera errónea que pueden repercutir de manera negativa en la calidad subjetiva final.
- Por su parte, las secuencias QCIF, presentan una clasificación más estable, pues apenas aparecen MB aislados que empeoren la consistencia espacial de la clasificación, como sucede en vídeos de tamaño superior. Además, existen MB pertenecientes a los objetos de interés que se clasifican como con mucho movimiento, cuando se está realizando un seguimiento del mismo, de modo que dificulta la delimitación de las regiones de interés, tal y como se puede observar en la Figura 4.22. Existe también el caso en que MB adyacentes a los

objetos de interés se clasifican como éstos, haciendo que la región de interés se expanda (véase Figuras 4.24 y 4.26).

Es necesario señalar que, en la mayoría de las secuencias, la influencia de las regiones uniformes no es tan destacable debido a la cantidad de información que incluyen los MB en vídeos de este tamaño. Así, si se comparan los vídeos clasificados "bohemia_morphoOFF_Thx.avi" y "bohemia6_qcif_morphoOFF_Thx.avi", para un umbral proporcional a la resolución del vídeo, se puede observar que la clasificación en el caso QCIF es más uniforme y el problema de las regiones de textura homogénea se ve reducido considerablemente.

- Por último, además de percibir inconsistencia espacial en los mapas binarios, si se visualiza la clasificación de cada secuencia completa, se puede apreciar inconsistencia temporal; algunos vídeos representativos pueden ser "football_morphoOFF_Thx.avi" y "bohemia_morphoOFF_Thx.avi". Ésta aparece en forma de parpadeos problemáticos a la hora de asignar el incremento de QP correspondiente y visualizar el resultado de la codificación.

El análisis subjetivo de la clasificación y las conclusiones extraídas al respecto parecen indicar la necesidad de una mejora en el mapa binario generado por el clasificador λ_1, que resuelva los casos citados anteriormente siempre que sea posible. Como consecuencia, surgen las etapas adicionales de mejoras detalladas en los siguientes capítulos.

Capítulo 5

Mejoras del algoritmo de clasificación de movimiento

5.1 Refinamiento del campo de vectores de movimiento

5.1.1 Introducción

En el capítulo anterior se han detallado ciertos casos anómalos detectados en la clasificación del movimiento, extraídos del análisis subjetivo de los resultados correspondientes a las pruebas de configuración y ajuste de parámetros del algoritmo EMJ. Además, el problema principal encontrado en la clasificación es la inconsistencia temporal, causada por las variaciones entre mapas binarios de planos consecutivos, y la inconsistencia espacial, causada por la no uniformidad de la clasificación del plano, debida a la presencia de huecos en regiones uniformes y MB aislados (*outliers*) en la clasificación. Esto podría repercutir de forma negativa sobre la calidad subjetiva del vídeo codificado, pues, si las asignaciones de QP a cada MB difieren mucho de un plano al siguiente, será perceptible un efecto parpadeo que degradará la calidad del vídeo.

Como consecuencia de todo ello, surge, en primer lugar, esta nueva etapa para subsanar los problemas citados en la medida de lo posible y conseguir, por tanto, una clasificación más ajustada al movimiento real de cada escena, así como una calidad subjetiva adecuada.

En concreto, esta etapa propuesta se encarga de la mejora del mapa de vectores generado por el algoritmo EMJ; previa a la clasificación, se trata, por tanto, de una etapa de post-procesado. Está destinada a cumplir las siguientes tareas: rellenar huecos (zonas clasificadas como poco movimiento en regiones de mucho movimiento), y eliminar MBs aislados clasificados como mucho movimiento. Para ello se considera adecuada la aplicación de operaciones morfológicas sobre el módulo del campo de vectores. Si la etapa morfológica propuesta resulta eficiente, la consistencia espacial del mapa binario se vería mejorada y, consecuentemente, la consistencia temporal de la clasificación también.

A continuación, se describen las herramientas que ofrece la morfología para elaborar la etapa de post-procesado, así como las operaciones definitivas seleccionadas y las pruebas realizadas al respecto.

5.1.2 Fundamentos de morfología matemática

La morfología matemática es una herramienta aplicada al análisis de imágenes digitales que se encarga del estudio de la forma o la estructura de los objetos presentes en el plano de interés y es comúnmente utilizado tras un proceso de segmentación.

El análisis morfológico de las imágenes permite extraer componentes útiles en la representación y descripción de la forma de las regiones (fronteras, esqueletos,…), y también, obtener características relevantes de los objetos de la imagen (parámetros de forma, tamaño,…). Por su parte, el procesado morfológico, pretende transformar la forma o la estructura de objetos presentes en la imagen. Existen tres tipos de morfología, en función del tipo de imagen sobre la que se trabaje: binaria, de niveles de gris y de imágenes en color; en este caso, hay que centrarse en morfología de niveles de gris, pues no se dispone de una imagen binaria, sino de valores discretos correspondientes al módulo de los vectores.

Los fundamentos del análisis y procesado morfológico se basan en el álgebra de conjuntos y en la topología, así en todo procesado de este tipo se encuentran tres elementos implicados:

(1). Conjuntos (Imágenes).
(2). Elementos Estructurantes (EE).
(3). Operadores Morfológicos (dilatación, erosión, apertura/cierre).

Toda operación morfológica utiliza un elemento estructurante (EE). Consiste en un patrón de ajuste para analizar la estructura geométrica de la imagen, denominado elemento estructurante (EE); éste se desplaza sobre la imagen y analiza su posición en relación al primer plano y al fondo de la misma. Se posiciona el centro del EE sobre cada píxel de la imagen original aplicando la operación morfológica sobre los puntos situados bajo él. La forma del EE puede ser cualquiera, pero los más usados son el cuadrado (o rectángulo) y el círculo.

0	1	0
1	1	1
0	1	0

1	1	1
1	1	1
1	1	1

Figura 5.1. Ejemplos de EE, con centro señalado en rojo.

Existen dos operaciones básicas a partir de los cuales se construye el resto de operaciones morfológicas: dilatación y erosión.

- **Dilatación:** La operación de dilatación consigue rellenar entrantes en los que no cabe el EE (rellena huecos). En el caso de imágenes en escala de grises, según detalla [27], el EE se sitúa en el punto de interés (por ejemplo, el MB señalado en gris en la Figura 5.2) y sustituye el valor de éste por el máximo valor entre aquellos valores sobre los que se sitúa el EE. En las Figuras 5.2 y 5.3, podemos observar que el EE cubre un cuadrado de tamaño 2x2 MBs, pues se trabaja a nivel de MB, no de píxel.

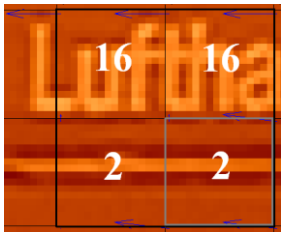


Figura 5.2. Antes de la dilatación.

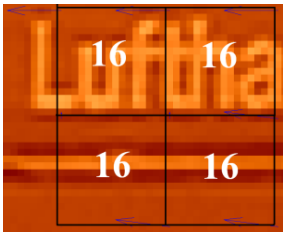


Figura 5.3. Después de la dilatación.

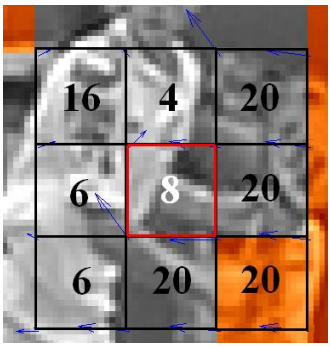


Figura 5.4. Antes de la erosión.

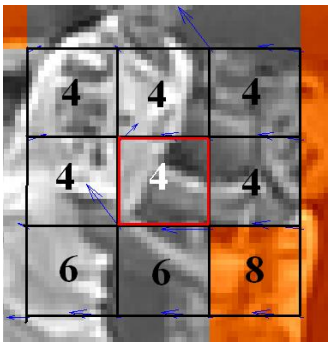


Figura 5.5. Después de la erosión.

- **Erosión:** La operación de erosión elimina elementos en los que no cabe el EE, y al contrario que la dilatación, en escala de grises sustituye el valor en punto de análisis (por ejemplo, el MB señalado en rojo) por el valor mínimo de aquellos MBs englobados por el EE, que en el caso de las Figuras 5.4 y 5.5, tiene un tamaño de 3x3 MBs.

Existen además otras operaciones morfológicas que se construyen a partir de las mencionadas anteriormente, los filtros morfológicos, que se describen a continuación.

- **Apertura:** La apertura consiste en una erosión seguida de una dilatación con el mismo elemento estructurante. Esta operación comúnmente se aplica para alisar contornos, eliminar protuberancias donde no quepa el EE o separar objetos en puntos estrechos (véase Figura 5.6).

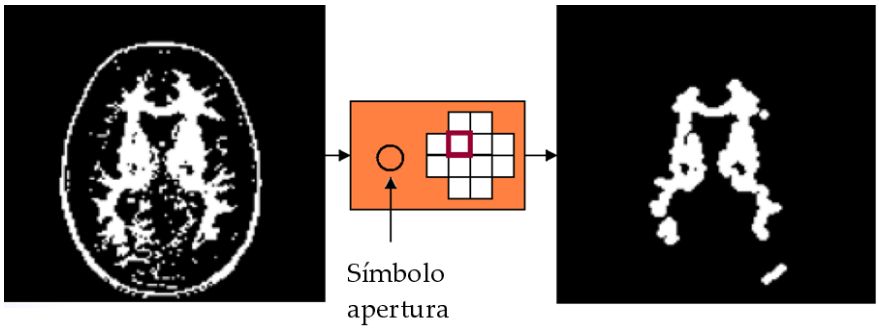


Figura 5.6. Ejemplo de apertura.

- **Cierre:** Por el contrario, el cierre consiste en una dilatación seguida de una erosión con el mismo elemento estructurante. Esta operación se suele aplicar para rellenar agujeros pequeños, eliminar entrantes

pequeños en un objeto o conectar objetos vecinos. La Figura 5.7 se corresponde con un ejemplo de relleno de huecos.

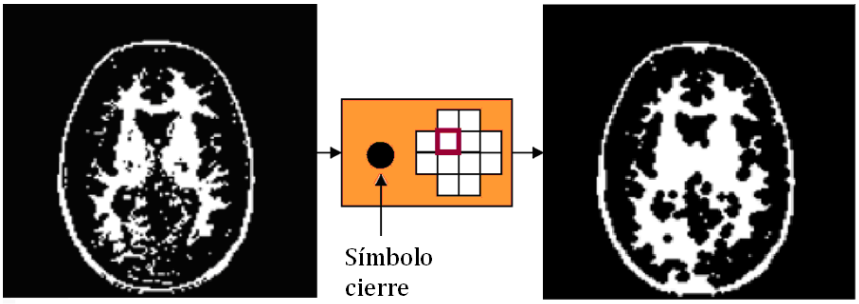


Figura 5.7. Ejemplo de cierre.

5.1.3 Etapa de post-procesado

Entre la salida del algoritmo EMJ y el clasificador de movimiento, se propone aplicar una etapa de post-procesado para conseguir una clasificación binaria más fidedigna al movimiento real de la escena, de modo que se rellenen los huecos que aparecen en las zonas homogéneas y se consiga delimitar aún más las zonas de mayor interés en cuanto a su movimiento característico. Para ello, se propone la aplicación de operaciones morfológicas, comúnmente utilizadas en el análisis de imágenes; se aplican sobre el módulo del campo de vectores de movimiento (VMs), pues es necesario recordar que el clasificador toma como entrada el módulo de los mismos.

Antes de pasar a evaluar el nuevo bloque de post-procesado, cabe hablar en primer lugar del coste computacional que esta etapa supone. Nótese que el algoritmo EMJ genera un mapa que representa el módulo de los VMs asociados a los MBs de un plano. Por consiguiente, el coste computacional de la etapa de post-procesado es bajo porque a su entrada se toman imágenes de MBs, no de píxeles. Así, si el tamaño del vídeo es CIF, la imagen de entrada a la etapa de post-procesado tendrá un tamaño de 22x18 unidades (352 columnas equivale a 22 MBs y 288 filas a 18 MBs). Si el tamaño del plano es SD con 720x576 píxeles, entonces se tomarán imágenes de 45x36 unidades.

El diagrama de bloques definitivo de la etapa se corresponde con la Figura 5.8. Se requiere por tanto de una combinación de dilatación y erosión, que

favorezca la delimitación de las regiones de poco movimiento con respecto a las de movimiento pronunciado:

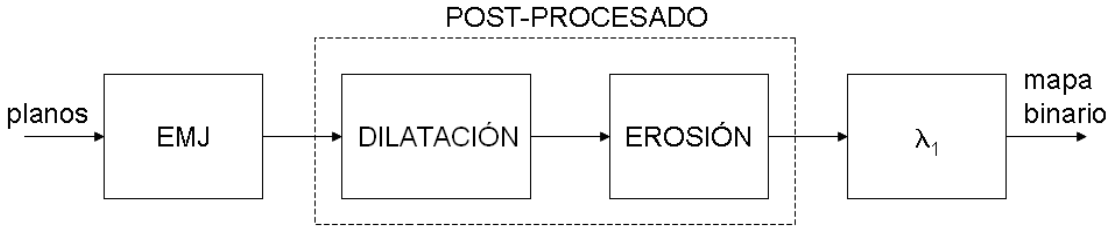


Figura 5.8. Etapa de post-procesado definitiva.

A pesar del aspecto final del diagrama de la etapa de post-procesado, ésta, inicialmente, estaba compuesta por otras dos operaciones conectadas en serie: cierre y erosión, debido a que el objetivo de cada operación va acorde con las necesidades a cubrir con esta etapa. Así, el algoritmo de cierre suele estar destinado a rellenar huecos (uno de los principales problemas a solventar), y por otro lado, la erosión tiene como meta la eliminación de esos MBs o pequeñas regiones aisladas que surgen en la clasificación. Por tanto, tras realizar una pequeña prueba de evaluación con la primera versión de la etapa de post-procesado añadida al sistema, se observó que la mejora obtenida era considerable y la salida del clasificador era más fiable al movimiento real de la escena y más consistente temporalmente.

La etapa de post-procesado se compone finalmente de una operación de dilatación seguida de una erosión, por cuestiones de coste computacional e implementación en el codificador. El cierre es una operación a realizar en dos pasos (dilatación+erosión) que no se puede simplificar. Y, si tras un cierre se aplica una erosión, ambas erosiones realizadas en total se pueden agrupar en una sola, de modo que en primer lugar se realice una dilatación con un EE característico y, después, una erosión con un EE diferente. Simplificando las operaciones morfológicas se consigue el mismo efecto que el obtenido con las erosiones concatenadas.

Tras realizar una batería de pruebas adecuada, se han preestablecido los valores de los elementos estructurantes de cada operación, en función del tamaño del vídeo, según recoge la Tabla 5.1. Como se puede observar, existe una relación proporcional en los EEs asignados a los distintos tamaños de vídeo; como es lógico, el tamaño del EE disminuye a medida que disminuye el tamaño del plano en la secuencia. Es necesario señalar que inicialmente, en el caso de las secuencias QCIF

se recurrió a una apertura morfológica con el fin de extender las regiones de interés, sin embargo las pruebas demostraron que esta operación no es capaz de solucionar este problema al completo, y que los valores definitivos de EE utilizados ofrecen una solución más adecuada.

En el siguiente apartado se recogen los resultados obtenidos en las pruebas de evaluación de esta etapa y se adjuntan las observaciones realizadas al respecto.

	QCIF	CIF	SD
Tamaño EE dilatación	▮	▮	▮
Tamaño EE erosión	▮	▮	▮

Tabla 5.2. Elementos estructurantes.

5.1.4 Análisis de los resultados con etapa de post-procesado

A continuación, se adjunta una muestra del efecto de la nueva etapa de post-procesado propuesta sobre la clasificación binaria obtenida. Las imágenes son representativas de las mejoras que introduce la morfología y también de los problemas que no puede resolver por completo. Las secuencias implicadas en la batería de pruebas son tanto de tamaño SD, como CIF y QCIF, para poder comparar las clasificaciones en cada uno de los casos y analizar los resultados correspondientes.

En primer lugar, en la clasificación correspondiente a la secuencia "*Bohemia*" se puede observar la mejora en la delimitación del fondo si se comparan las Figuras 5.9 y 5.10, pues, se observa la desaparición de ciertos huecos en el fondo al activar la etapa de post-procesado. Sin embargo, en las zonas correspondientes a la pared, debido a su homogeneidad y a su tamaño, no se consigue rellenar por completo; también se puede apreciar que las siluetas de los objetos de interés quedan mejor delimitadas. En general, se consigue una mejora de la representación de la escena en la secuencia completa, disminuyendo los parpadeos producidos por la inconsistencia espacio-temporal presente (véase "bohemia_morphoON_Thx.avi", "football_morphoON_Thx.avi" disponibles en el DVD adjunto).

Figura 5.9. Salida clasificador $\lambda_1 =$ ■ con post-procesado. Secuencia "Bohemia" (352x288).

Figura 5.10. Salida clasificador $\lambda_1 =$ ■ sin post-procesado. Secuencia "Bohemia" (352x288).

Por otro lado, las Figuras 5.11 y 5.12 son representativas de un aspecto negativo de la nueva etapa. Consiste en que la morfología no puede solventar en todos los casos el problema de eliminación de MBs aislados, aunque sí puede conseguir una reducción de los mismos.

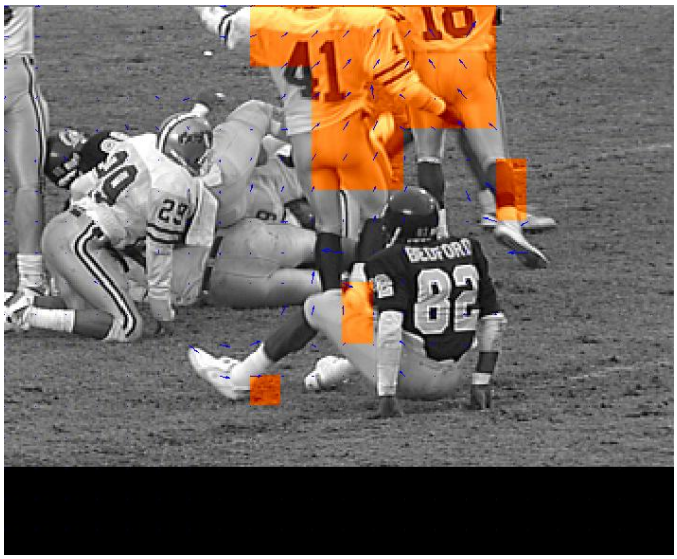


Figura 5.11. Salida clasificador $\lambda_1 =$ ■ con post-procesado. Secuencia "Football" (352x288)

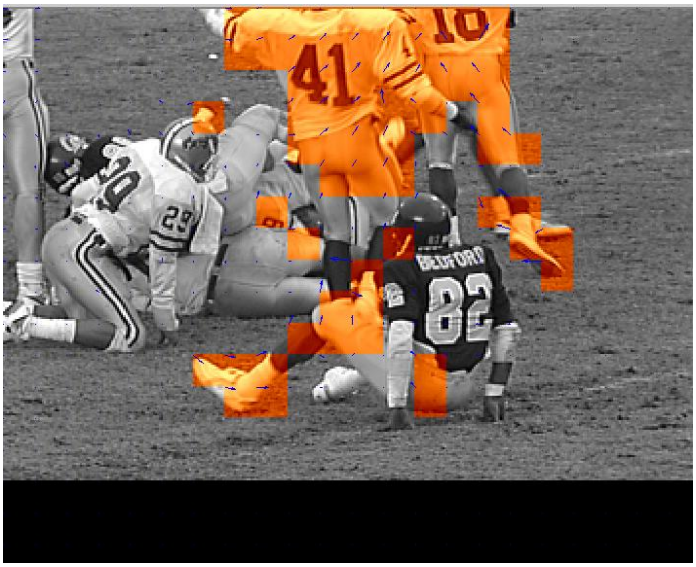


Figura 5.12. Salida clasificador $\lambda_1 =$ ■ sin post-procesado. Secuencia "Football" (352x288).

En contraposición a esta muestra negativa de la eficiencia de la etapa, se incluyen las siguientes capturas, relativas a la eliminación de MBs aislados también. Estas figuras demuestran la mejora que aporta la morfología en otras secuencias, como es el caso de "*Ultimo Samurai*"; se observa la desaparición de MBs clasificados como con mucho movimiento que deterioraban la consistencia

del clasificador binario de movimiento, así como su correspondencia con el movimiento real de la escena.

Figura 5.13. Salida clasificador λ_1=[redacted] con post-procesado. Secuencia "Último Samurai" (720x576).

Figura 5.14. Salida clasificador λ_1 =[redacted] sin post-procesado. Secuencia "Último Samurai" (720x576).

Además, la morfología no sólo se enfrenta al problema de los MBs aislados, pues, en "*Bohemia*" aunque tienen más peso las mejoras conseguidas, las zonas homogéneas no consiguen rellenarse por completo, situación común en otras secuencias como "*LOTR*". En la Figura 5.15 se observa que no sólo el gorro no es

clasificado como (1), sino que los bordes correspondientes al mismo que antes lo estaban, ahora han invertido su clasificación.

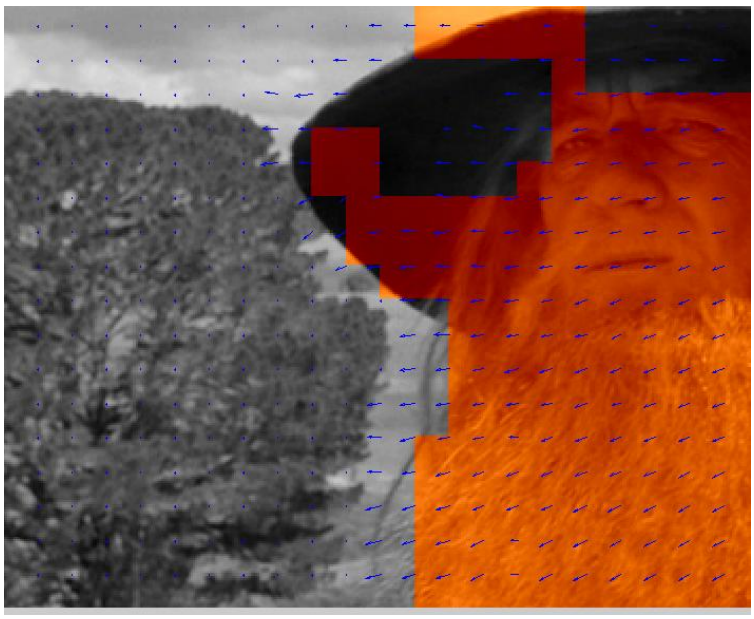


Figura 5.15. Salida clasificador λ_1 = [redacted] con post-procesado. Secuencia "LOTR" (352x288).

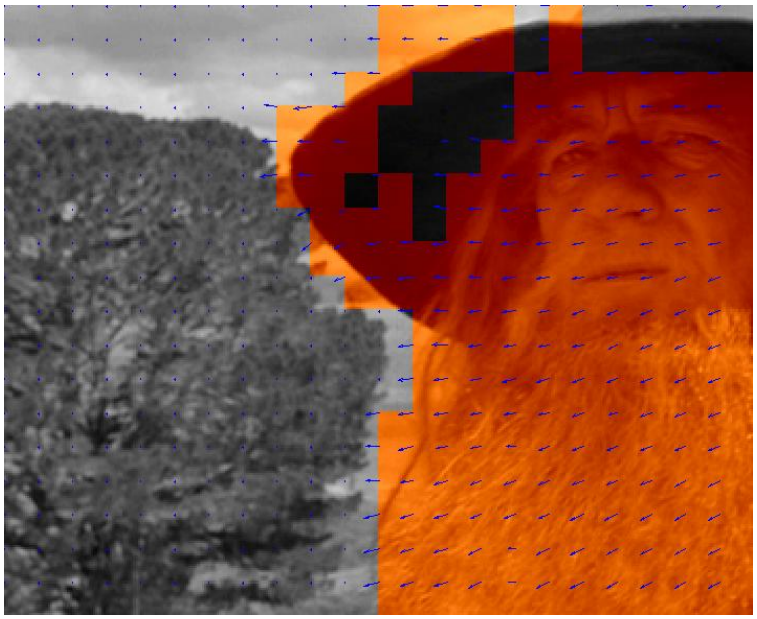
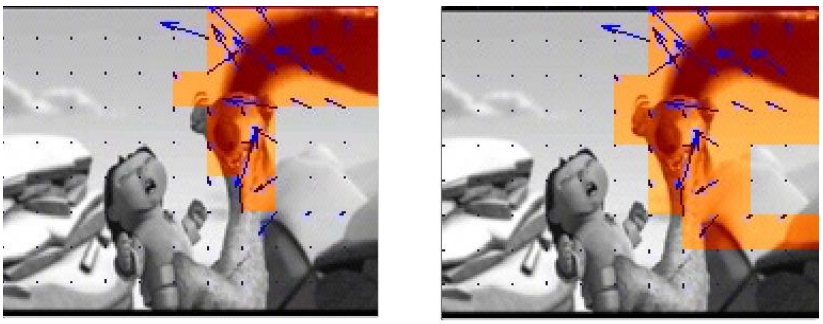


Figura 5.16. Salida clasificador λ_1 = [redacted] sin post-procesado. Secuencia "LOTR" (352x288).

Una posible solución para rellenar huecos grandes podría ser aumentar el tamaño del EE en la dilatación, pero esto tendría consecuencias negativas en otros

aspectos, porque ciertas regiones cuyo movimiento es nulo serían clasificadas como con mucho movimiento. Así, si se aumentara el valor del EE sobre la secuencia "*Bohemia*", no sólo se rellenarían los huecos relativos a la pared, sino que las siluetas de las regiones de interés también lo harían, careciendo de significado el mapa binario resultante. Por lo tanto, parece más sensato descartar esta posibilidad puesto que el hecho de que regiones uniformes no queden rellenas completamente no repercute en la calidad de la secuencia, pues, aunque el valor de QP asignado a estas áreas difiera de los MBs adyacentes, éstas se codificarían sin distorsión apreciable.

Con respecto a las secuencias QCIF, las mejoras para vídeos de este tamaño son semejantes a las citadas, y, adicionalmente, existen un par de observaciones a destacar. La primera de ellas, relacionada con una situación anómala encontrada en vídeos QCIF en el apartado 4.2, consiste en que los MBs alrededor de las regiones de interés con movimiento elevado que se incluían en la clasificación, ahora se eliminan de manera parcial o completa en ciertos casos, según la secuencia en estudio (véase Figura 5.17). Otra mejora conseguida gracias a la morfología es la delimitación de las regiones de interés, eliminando aquellos MBs clasificados a (1) dentro del objeto de interés cuando se está realizando un seguimiento del mismo; un ejemplo de este caso estaría representado por la Figura 5.18 de la secuencia "bus_qcif_morphoON_Thx.avi" disponible en el DVD adjunto, donde se consiguen eliminar ciertos MBs del interior del autobús que han sido clasificados de manera errónea haciendo que el autobús quede más delimitado que en la versión anterior del sistema.



(a) (b)

Figura 5.17. Salida clasificador λ_1 = [redacted]. Con post-procesado (a). Sin post-procesado (b). Secuencia "Ice Age" (176x144).

(a) (b)

Figura 5.18. Salida clasificador λ_1 = [redacted]. Con post-procesado (a) y sin post-procesado (b). Secuencia "Bus" (176x144).

5.1.5 Conclusiones

La etapa propuesta en este apartado conformada por un procesado morfológico a la salida del algoritmo EMJ, resulta beneficiosa para el mapa binario característico de la clasificación. Este módulo consigue solventar en la medida de lo posible los casos anómalos que presenta la salida del clasificador λ_1, aunque no es efectivo en todos los tipos de secuencias. Mediante operaciones morfológicas sencillas permite reducir el número de MBs aislados con movimiento pronunciado y rellenar huecos en zonas con movimiento. Además, hay que destacar la mejora en la consistencia espacio-temporal de la clasificación obtenida en la mayoría de los casos, hecho favorable desde el punto de vista subjetivo.

Por último, es necesario señalar que la etapa es computacionalmente sencilla, pues las dos operaciones realizadas sobre cada plano de la secuencia, trabajan a nivel de MB, no de píxel; y, la etapa proporciona un mapa binario de clasificación mejorado independientemente del tamaño de vídeo.

5.2 Procesado previo de la secuencia de entrada

5.2.1 Introducción

Uno de los requisitos más importantes que debe cumplir un plano para que el movimiento de los objetos con respecto al de cámara pueda ser detectado fielmente por el algoritmo EMJ, es la información de bordes o de alta frecuencia de los mismos, pues cuánto más pronunciada sea, más factible resulta la detección del movimiento. Por ello, en pruebas realizadas en módulos anteriores se han encontrado inconvenientes en regiones de textura uniforme que presentan movimiento; el sistema no es capaz de detectarlo, a menos que se active la etapa de post-procesado, que mejora notablemente la clasificación del movimiento, aunque no resuelve del todo la situación.

Cabe pensar en la aplicación de otra etapa, esta vez previa a EMJ, que se encargue de realzar la información de bordes, incluso en aquellas secuencias donde esta información apenas es visible a simple vista. Un ejemplo muy representativo es una secuencia poco contrastada (como el caso de la secuencia de prueba "*África*", disponible en el DVD adjunto), que se caracteriza por un histograma de luminancia muy concentrado en un rango pequeño en la escala de grises, de modo que la información de alta frecuencia está escondida en los niveles de grises y, por tanto EMJ no actúa correctamente al no disponer de un grado de detalle suficiente. Otro caso en el que se requiere un realce de bordes es una secuencia en donde la textura es pseudo-uniforme, como puede ser el asfalto de una carretera en vídeos como "*Highway*" o "*Corvette*" (vídeo característico incluido en el DVD), utilizados en la evaluación de la etapa también.

En definitiva, con la etapa de pre-procesado se espera una mejora en la clasificación en un grupo de secuencias similar a alguno de los dos casos anteriormente descritos, y sin que en el resto de secuencias el resultado de la clasificación empeore.

Para acometer esta tarea, se propusieron diversas técnicas de realzado de bordes, que se describen a continuación; entre ellas se selecciona una definitiva que conforma el módulo de pre-procesado que se propone en este apartado.

5.2.2 Técnicas de realzado de bordes

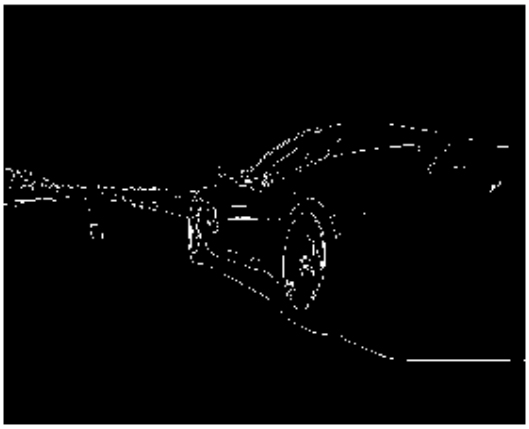
Varios métodos de realzado de bordes se propusieron para conformar la etapa de pre-procesado. Con el fin de determinar el más apropiado para cubrir las necesidades planteadas, las técnicas fueron evaluadas con una serie de vídeos: "*África*", "*Corvette*" y "*Airshow*", de los cuales se adjuntan capturas representativas. Estos vídeos presentan ciertas irregularidades, como la falta de contraste o la aparición de detalle poco definido, que no permiten que el algoritmo EMJ actúe eficientemente y pueda generar un mapa de vectores acorde con el movimiento real presente en la escena. En estas secuencias, el objeto de interés no queda bien identificado a la salida del clasificador debido al movimiento del fondo y las altas frecuencias escondidas en éste.

- Filtrado paso alto

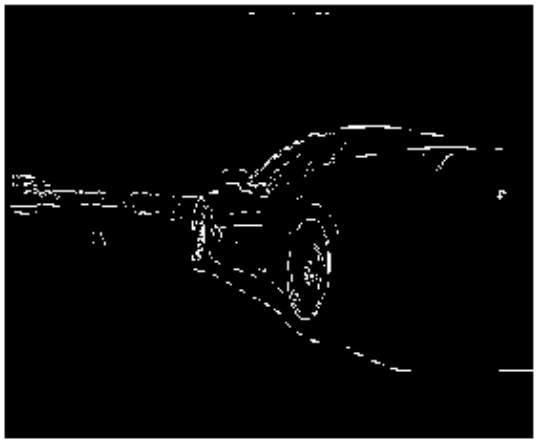
  La primera solución planteada fue el filtrado paso alto de los planos y el empleo del resultado de los mismos como entradas a EMJ. El uso del filtrado paso alto se debe a que el fin de éste es conseguir una imagen exclusivamente con información de borde, que es la que requiere precisamente EMJ para obtener el movimiento. Se probaron varias máscaras paso alto: filtro de *Roberts*, filtro de *Sobel*, filtro de *Prewitt* y filtro de *Canny*. Una muestra de los resultados obtenidos se puede observar en la Figura 5.19; las tres primeras técnicas no proporcionan imágenes paso alto muy favorables porque las máscaras son muy sencillas, de modo que no son capaces de detectar toda la información de alta frecuencia esperada como, por ejemplo, la textura granular de un asfalto en la secuencia "*Corvette*". En cambio, con el filtro de Canny se consigue extraer la información esperada: una imagen binaria con todo el contenido de alta frecuencia (véase Figura 5.19. e).

(a) Plano original

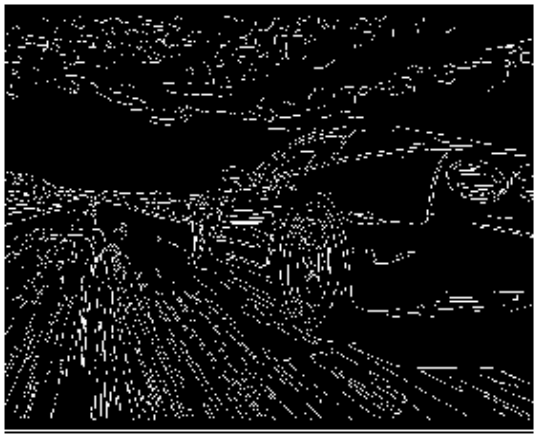
(b) Sobel



(c) Prewitt



(d) Roberts



(e) Canny

Figura 5.19. Filtrado paso alto. Secuencia "Corvette" (704x576)

Por lo tanto, el siguiente paso a realizar consistió en aplicar el filtro de Canny sobre los planos de la secuencia y, posteriormente, ejecutar el algoritmo EMJ para observar el mapa de vectores resultante. Desgraciadamente, los resultados no fueron nada buenos, puesto que la función de coste de EMJ está optimizada para imágenes en escala de grises, no binarias; entonces, no existía ninguna información de luminancia que generara valores de MAD decisivos en la elección del coste mínimo. En cualquier caso, finalmente se desestimó el uso del filtro de Canny porque computacionalmente es muy costoso, pues se fundamenta en la primera derivada y no recurre únicamente a una máscara, y, además, porque habría que reajustar los coeficientes de la función de coste en EMJ.

- Reforzamiento de bordes

Mediante el uso de ciertas máscaras predeterminadas se refuerzan las altas frecuencias dejando intactas las bajas en una imagen, y se aumenta la diferencia de amplitudes entre los píxeles que forman el borde. Por tanto, parece lógico recurrir a ellas para comprobar si resultan eficaces en esta tarea, y así se hizo.

La máscara utilizada en las pruebas fue seleccionada al azar, puesto que el objetivo de todas las posibles es el mismo y se utilizaron secuencias con texturas problemáticas diferentes sobre las que poder reflejar su efecto. La máscara fue la siguiente:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Sin embargo, los resultados obtenidos a la salida de EMJ, aplicando el conjunto de vídeos de evaluación, reflejaron que este tipo de filtrado no aporta ninguna mejora con respecto a los resultados que proporciona el sistema con solamente la etapa de post-procesado activada.

- Aumento del contraste de la imagen

  Sabiendo que un borde es más prominente cuando la diferencia de luminancia es grande entre el píxel contenido en el borde y el píxel fuera del mismo, puede ser factible el empleo de alguna técnica de aumento del contraste de la imagen para realzar la presencia de alta frecuencia. En concreto, se propone la técnica de igualación o ecualización del histograma, que consiste en la obtención de una nueva imagen con una distribución uniforme de niveles.

  Con objeto de llevar a cabo una primera evaluación sobre la validez de esta técnica, se hicieron pruebas con la secuencia "*Corvette*". En este vídeo, la información de movimiento que interesa es la relativa a la carretera, sin embargo, cuando se aplica solamente el algoritmo EMJ seguido de post-procesado, este movimiento no se detecta. Al incluir una etapa previa de ecualización del histograma, se observa que la información de movimiento en el asfalto ha aparecido, debido a que las altas frecuencias escondidas en él han sido realzadas considerablemente por el aumento del contraste.

  Por otra parte, interesa que la aplicación de la etapa de pre-procesado no empeore los resultados de clasificación en otras secuencias ya utilizadas en las evaluaciones de las etapas EMJ y post-procesado; por ese motivo, también se evaluó el sistema con la secuencia "*Bohemia*". La clasificación obtenida en este caso es similar a la conseguida sin la etapa de pre-procesado activada, pero con la ecualización se consigue reflejar mejor el movimiento de la cámara en ciertas regiones de la fachada en "*Bohemia*", así como en el pavimento en el caso de "*Corvette*".

En definitiva, tras evaluar todas las técnicas de realzado disponibles la que ofrece mejores resultados es la de ecualización del histograma, por ello, a continuación se realiza una descripción teórica al respecto y, después, se analizan los resultados de las pruebas de evaluación pertinentes.

5.2.3 Ecualización del histograma

Después de aplicar cada una de las técnicas anteriormente detalladas y analizar los resultados obtenidos, se concluye que la ecualización del histograma proporciona los mejores resultados que el resto de técnicas evaluadas. Estos resultados demuestran que el aumento del contraste de la imagen mejora la detección del movimiento. Por esta razón, se decide implementar en el codificador H.264/AVC el ecualizador de histograma para evaluar y comprobar que la mejora en la clasificación es notable con respecto a la estimación de movimiento sin pre-procesado. El conjunto de vídeos empleado fue mayor al utilizado en baterías de pruebas anteriores y los resultados se recogen más adelante.

5.2.3.1 Fundamentos teóricos

El histograma de una imagen se define como la información de probabilidad de aparición de las distintas tonalidades de color que se pueden dar en cada caso, ya sean los distintos tipos de colores para imágenes en color o la escala de grises en caso de imágenes monocromáticas. En cualquier caso, el histograma proporciona una descripción de la apariencia global de la imagen, indicando los niveles en los que se concentra la imagen.

La ecualización o igualación del histograma es una técnica que permite mejorar el contraste de luminancia en una imagen, gracias a la uniformidad que se consigue en el histograma de la imagen de salida; esta uniformidad se refiere a disponer de una misma frecuencia de aparición en cada uno de los niveles de gris del histograma, ampliando así el rango de luminancia de la imagen.

Las mejoras relativas a su apariencia visual que se pueden conseguir en la imagen, tras la aplicación de esta ecualización, no son centro de atención para la consecución de nuestro objetivo en el desarrollo de esta etapa de pre-procesado, pero bien es cierto que van ligadas a ese reforzamiento de los bordes que se consigue y a esa mayor utilización de los recursos disponibles en la imagen.

También es necesario señalar ciertos inconvenientes que pueden surgir al recurrir a esta técnica como pueden ser:

- *Pérdida de la información***.** Como consecuencia de modificar, por ejemplo, un grupo de píxeles con diferentes niveles y asignarles un mismo nivel a todos.

- *Realce de algún error indeseado*. Si la secuencia de entrada al sistema ha sido previamente codificada a calidad alta y decodificada, el pre-procesado puede potenciar la aparición del ruido de codificación generado.

5.2.3.2 Algoritmo de ecualización del histograma

A continuación, tras conocer las características de esta técnica así como sus mejoras e inconvenientes, se describe el algoritmo implementado. El correspondiente algoritmo de ecualización del histograma está compuesto por las siguientes etapas:

(a). Histograma del plano.

La generación del histograma de cada plano de la secuencia correspondiente consiste en la creación de un vector de 256 elementos, correspondientes con las diferentes tonalidades de gris que tiene la imagen (suponiendo una resolución en amplitud de 8 bits/píxel), que contiene las frecuencias de aparición de cada nivel entre todos los píxeles del plano. Para generarlo, se recorre cada uno de los píxeles del plano y se aumenta una unidad el valor almacenado en la posición de dicho array, que se corresponde al nivel de gris del píxel actual que está siendo analizado. Una vez que se dispone de las frecuencias de aparición se divide cada una de ellas por el número total de píxeles del plano, obteniendo la frecuencia relativa,

$$p_u(u_i) = \frac{n(u_i)}{n} \qquad (19)$$

donde $n(u_i)$ denota el número de píxeles que tienen un nivel u_i determinado y n es el número total de píxeles.

En las Figuras 5.20 y 5.21 se puede observar un plano perteneciente al vídeo "*África*" y su respectivo histograma de frecuencias relativas a modo de ejemplo. En el histograma se aprecia que el rango de niveles de gris es estrecho, y aparece un nivel, en torno al valor 65, que concentra gran parte de la frecuencia de aparición. Por lo tanto, este plano está poco contrastado, y, por ello, el movimiento característico no es detectado por el algoritmo de estimación utilizado.

Figura 5.20. Plano en escala de gris. Secuencia "África" (656x544).

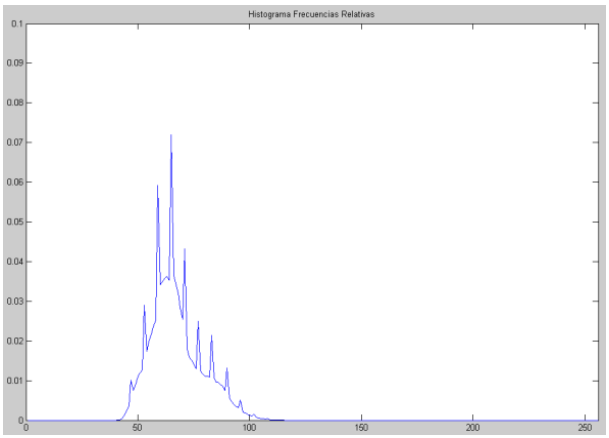


Figura 5.21. Histograma frecuencias relativas. Secuencia "África" (656x544).

(b). Histograma acumulado.

Este paso se corresponde con la suma acumulada de las frecuencias relativas calculadas previamente. Responde a la siguiente ecuación.

$$s_i = \sum_{j=0}^{i} p_u(u_j) = \sum_{j=0}^{i} \frac{n(u_j)}{n} \qquad (20)$$

A continuación se incluye el histograma acumulado correspondiente al plano de ejemplo de la secuencia "*África*".

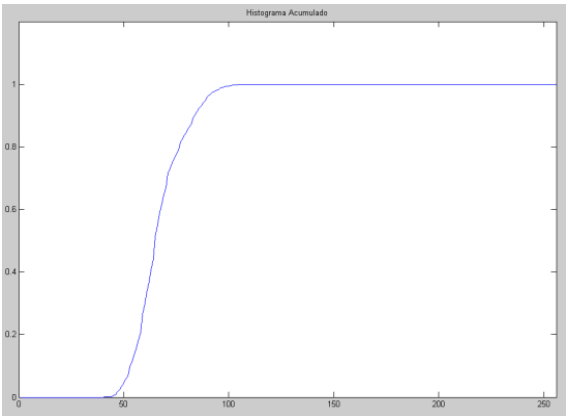


Figura 5.22. Histograma acumulado. "África" 656x544.

(c). Recuantificación.

El último paso a realizar en el proceso de ecualización del histograma consiste en reasignar los nuevos niveles de gris, tarea que se realiza aplicando la

expresión (21). El vector resultante se puede considerar como una LUT (*Look-Up Table*), encargada de asignar los nuevos niveles de gris de cada píxel.

$$s^* = Ent\left[\frac{s - s_{min}}{1 - s_{min}} \cdot 255 + 0.5\right] \qquad (21)$$

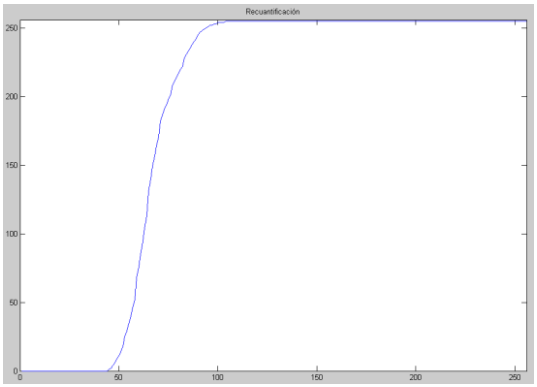


Figura 5.23. LUT para reasignación de niveles.

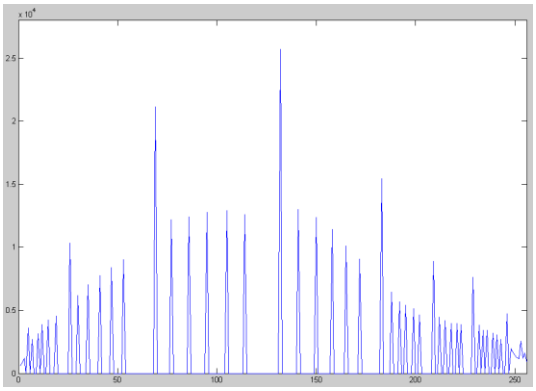


Figura 5.24. Histograma absoluto de plano ecualizado.

Figura 5.25. Plano ecualizado.

En la Figura 5.24 se puede observar la expansión que sufre el histograma del plano, haciéndolo más uniforme y, como consecuencia, un plano más contrastado, tal y como se percibe en la Figura 5.25. Además, si se compara esta última figura con la Figura 5.20, se identifica claramente el afloramiento de la información de detalle, destacado en la hierba y las ramas del fondo.

5.2.3.3 Análisis de los resultados

Con el fin de evaluar la eficiencia de la nueva etapa en el sistema se realiza una batería de pruebas. En el conjunto de vídeos seleccionado se incluyen varias secuencias problemáticas en pruebas anteriores debido a las características de

textura que presentan. La ecualización del histograma debe mejorar el mapa de vectores estimado así como la clasificación de esos vídeos peculiares, y, a su vez, no empeorar los resultados en el resto de secuencias a estudiar. A continuación se listan los vídeos utilizados, de entre los cuales, "*África*" y "*Corvette*" son las secuencias más conflictivas:

- Secuencias CIF: "*Bus*", "*Ice*", "*Pedestrian*", "*Football*", "*Stefan*".
- Secuencias SD: "*África*", "*Airshow1*", "*Airshow3*", "*Corvette*", "*Bohemia*", "*Ice Age*".

Todas estas secuencias han sido utilizadas previamente en pruebas de evaluación anteriores, en concreto, en las correspondientes a la etapa de post-procesado. Al disponer de los resultados obtenidos en esta batería de pruebas anterior, se recurrirá a ellos para realizar una comparativa con los obtenidos en estas pruebas, de modo que se pueda determinar la validez o no de esta etapa propuesta, así como la necesidad de una reconfiguración de los parámetros relativos a la morfología y/o el umbral de clasificación.

En primer lugar, se adjuntan capturas significativas de los efectos de la ecualización en las secuencias CIF de prueba (Figuras de la 5.26 a la 5.33), utilizando un umbral de clasificación de [redacted] píxeles. Para cualquier resolución de vídeo, a la izquierda se presentan los resultados obtenidos únicamente con la etapa de post-procesado activada, y a la derecha, la salida del clasificador con la etapa de pre-procesado habilitada junto con la de post-procesado.

Figura 5.26. Salida clasificador λ_1 = [redacted] sin pre-procesado. Secuencia "Bus" (352x288).

Figura 5.27. Salida clasificador λ_1 = [redacted] con pre-procesado. Secuencia "Bus" (352x288).

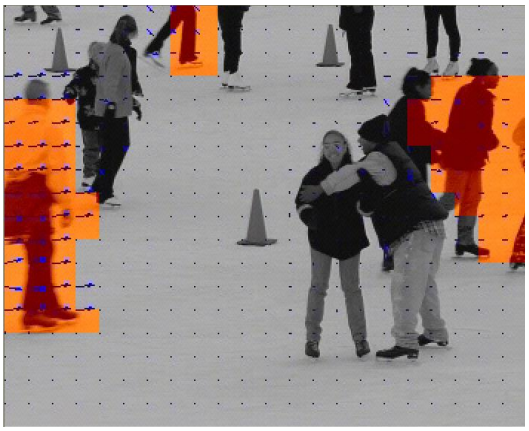


Figura 5.28. Salida clasificador $\lambda_1 =$ [redacted] sin pre-procesado. Secuencia "Ice" (352x288).

Figura 5.29. Salida clasificador $\lambda_1 =$ [redacted] con pre-procesado. Secuencia "Ice" (352x288).

Figura 5.30. Salida clasificador $\lambda_1 =$ [redacted] sin pre-procesado. Secuencia "Pedestrian" (352x288).

Figura 5.31. Salida clasificador $\lambda_1 =$ [redacted] con pre-procesado. Secuencia "Pedestrian" (352x288).

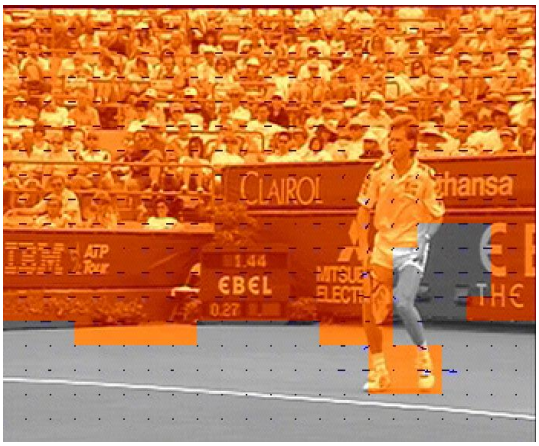


Figura 5.32. Salida clasificador $\lambda_1 =$ [redacted] sin pre-procesado. Secuencia "Stefan" (352x288).

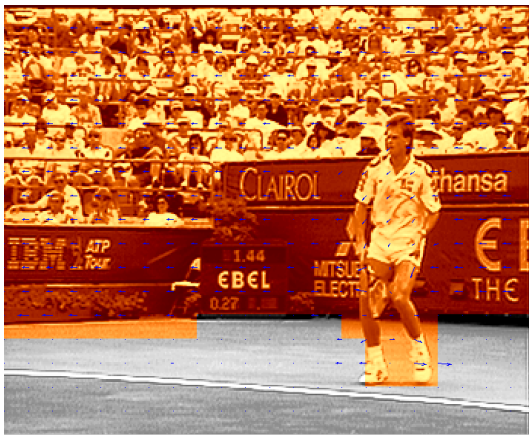


Figura 5.33. Salida clasificador $\lambda_1 =$ [redacted] con pre-procesado. Secuencia "Stefan" (352x288).

Por un lado, es necesario destacar que la nueva etapa permite rellenar huecos que la morfología no puede por sí sola. Así, esto se puede apreciar en la Figura 5.27, correspondiente a la secuencia "*Bus*", y en la Figura 5.33 de "*Stefan*". Las diferencias en cuanto a la cantidad de huecos cubiertos no son elevadas y, en general, los resultados obtenidos son similares en ambas versiones; sin embargo, la leve mejora es ventajosa.

Otra característica destacada de los resultados es que se consigue delimitar mejor la región de interés en ciertas secuencias, como por ejemplo, en "*Pedestrian*". Si se observan los planos correspondientes a este vídeo (Figuras 5.30 y 5.31), se aprecia la mejora a la hora de delimitar las siluetas cuando la ecualización está activada. Adicionalmente, si se visualiza la secuencia completa clasificada, se percibe una mayor consistencia temporal en la clasificación resultante.

Los resultados correspondientes a la secuencia "*Football*" no han sido añadidos, puesto que no aportan ninguna observación de interés con respecto a las mejoras conseguidas o a los errores introducidos con la nueva etapa. Por su parte, la secuencia "*Ice*" muestra ciertas peculiaridades introducidas por la ecualización. Esta operación ha hecho aflorar ruido de alta frecuencia en la zona de la pista de hielo, lo cual puede resultar positivo desde el punto de vista de la clasificación del movimiento, en caso de que el movimiento de cámara no fuera nulo. Sin embargo, también se puede observar cómo la ecualización hace que se oscurezcan mucho las siluetas de los patinadores, debido a que en el plano predomina el blanco; esto puede dar lugar a que el movimiento en el interior de objetos no sea detectado en secuencias con resoluciones mayores, pues esta secuencia en concreto no presenta este problema.

Tras realizar todas las observaciones anteriores sobre los resultados en secuencias CIF, se continúa con el conjunto de secuencias SD, de las cuales se muestran los siguientes planos representativos (Figuras de 5.34 a 5.45); el umbral de clasificación utilizado es de ■ píxeles.

Figura 5.34. Salida clasificador λ_1 = sin pre-procesado. Secuencia "África" (656x544).

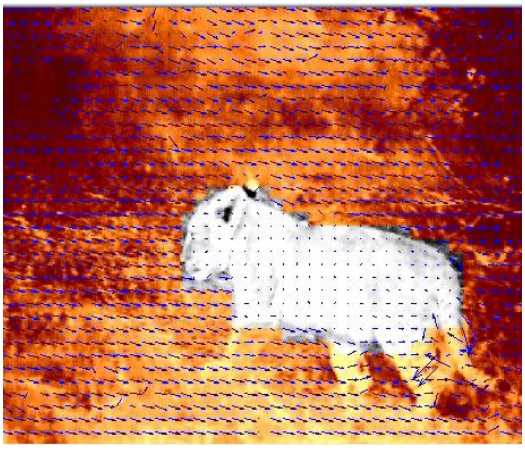


Figura 5.35. Salida clasificador λ_1 = con pre-procesado. Secuencia "África" (656x544).

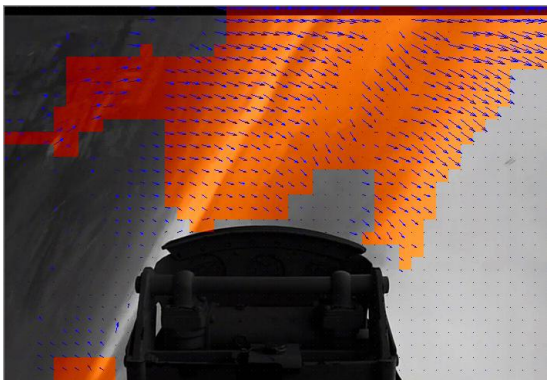


Figura 5.36. Salida clasificador λ_1 = sin pre-procesado. Secuencia "Airshow1" (704x480).

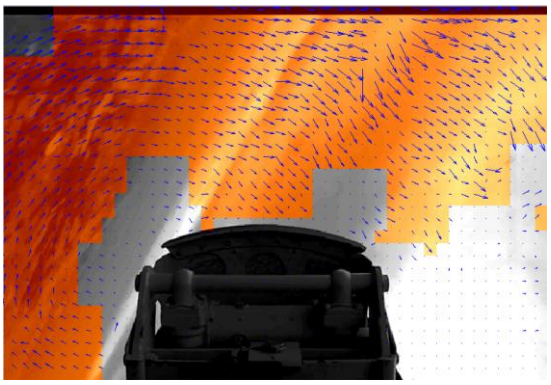


Figura 5.37. Salida clasificador λ_1 = con pre-procesado. Secuencia "Airshow1" (704x480).

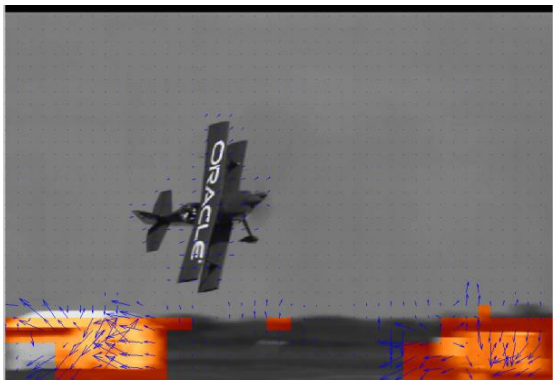


Figura 5.38. Salida clasificador λ_1 = sin pre-procesado. Secuencia "Airshow3" (704x480).

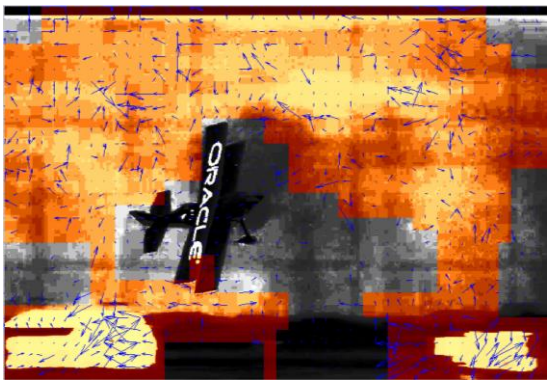


Figura 5.39. Salida clasificador λ_1 = con pre-procesado. Secuencia "Airshow3" (704x480).

Figura 5.40. Salida clasificador $\lambda_1 =$ [redacted] sin pre-procesado. Secuencia "Bohemia" (704x576).

Figura 5.41. Salida clasificador $\lambda_1 =$ [redacted] con pre-procesado. Secuencia "Bohemia" (704x576).

Figura 5.42. Salida clasificador $\lambda_1 =$ [redacted] sin pre-procesado. Secuencia "Corvette" (704x576).

Figura 5.43. Salida clasificador $\lambda_1 =$ [redacted] con pre-procesado. Secuencia "Corvette" (704x576).

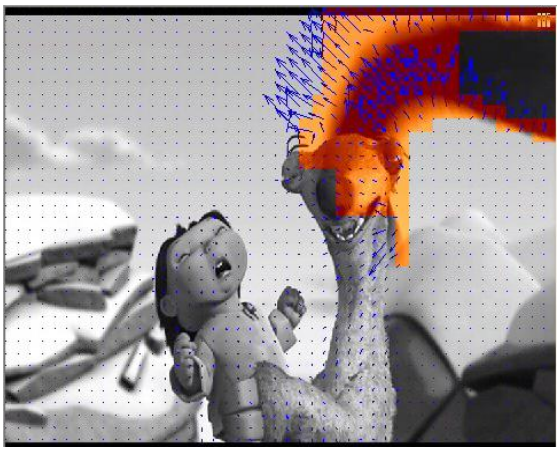


Figura 5.44. Salida clasificador $\lambda_1 =$ [redacted] sin pre-procesado. Secuencia "Ice Age" (720x576).

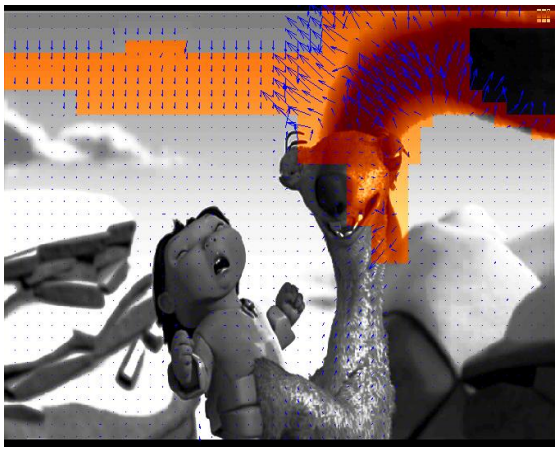


Figura 5.45. Salida clasificador $\lambda_1 =$ [redacted] con pre-procesado. Secuencia "Ice Age" (720x576).

En el caso de este conjunto de secuencias, el grado de detalle conseguido gracias a la ecualización es considerable y claramente observable. Al disponer de mayor detalle en los planos, el algoritmo EMJ proporciona un mapa de vectores más fiable, aunque esto no se cumple en todos los casos; pues, en las secuencias "*Airshow3*" y "*Ice Age*" se puede observar cómo la ecualización hace que el ruido de codificación se potencie en las regiones uniformes del plano (todas las secuencias SD no son originales, son reconstrucciones de bitstreams codificados a calidad alta). Así, en la Figura 5.45 se detecta movimiento en el fondo que realmente no existe; del mismo modo, en la Figura 5.39 se puede apreciar que, a pesar de que la clasificación parece esbozar una segmentación de las regiones de interés, los vectores del fondo que surgen como consecuencia de la ecualización son caóticos e infieles al movimiento real de la escena.

En contraposición al problema comentado, el resto de capturas muestran una mejora de la clasificación cuando la etapa de pre-procesado está activada. Así, en las capturas correspondientes a "*Bohemia*" y "*Airshow1*" se puede observar como el movimiento del fondo queda más diferenciado en regiones de textura uniforme; si se visualizan las secuencias correspondientes del DVD adjunto se puede apreciar dicha mejora ("bohemia_eq_morphoON_Thx.avi" y "airshow1_eq_morphoON_Thx.avi"), y además, en el caso de "*Bohemia*" la delimitación de las regiones de interés se ve favorecida con la activación de esta nueva etapa.

Por otro lado, con respecto a aquellas secuencias consideradas problemáticas en etapas anteriores, es necesario señalar que, en primer lugar, en "*Corvette*" (véase "corvette_eq_morphoON_Thx.avi") al potenciar las altas frecuencias, aparece el detalle de la calzada que permite detectar mejor el movimiento presente en la secuencia. Por su parte, el vídeo "*África*" presentaba problemas en la detección de movimiento cuando únicamente se contaba con la etapa de post-procesado, ahora, la ecualización realza los detalles del fondo, permitiendo al algoritmo EMJ detectar el movimiento en gran medida (véase "africa_eq_morphoON_Thx.avi"); así, con esta nueva etapa el seguimiento del objeto de interés queda visible mientras en la versión anterior, la clasificación apenas tenía sentido.

5.2.3.4 Conclusiones

En este apartado se ha realizado una breve explicación teórica de la funcionalidad propuesta, que conforma una etapa de procesado previa al algoritmo EMJ y que pretende solventar problemas que la etapa de post-procesado no consigue resolver en todos los casos. En concreto, esta nueva etapa recurre a la ecualización de los planos para realzar las altas frecuencias correspondientes a zonas pseudo-homogéneas con el objetivo de detectar mejor el movimiento y conseguir un mapa de vectores eficiente frente a estas regiones peculiares.

Tras realizar la batería de pruebas correspondiente y analizar los resultados obtenidos, se puede concluir que no se ha conseguido una solución eficiente para todas las secuencias de prueba; por lo tanto, existen ciertas secuencias que no toleran la ecualización, haciendo que esta etapa no sea genérica. Para mejorar la etapa de pre-procesado se proponen las siguientes soluciones:

- Un clasificador de movimiento que disponga de información temporal.
  Las inconsistencias temporales encontradas en los resultados de la mayor parte de las secuencias de prueba utilizadas hasta el momento, se deben en parte a la simplicidad del clasificador utilizado.
  Existe un único umbral prefijado que establece cuándo el módulo de un vector de movimiento representa un movimiento elevado o pequeño; por lo tanto, este clasificador no es robusto frente a posibles ruidos en la detección del movimiento, como puede ser la uniformidad del movimiento de cámara en alguna secuencia, como es el caso de "*Bohemia*". La introducción de información temporal en el clasificador binario λ_1 se trata en un apartado posterior, en el que se define la solución propuesta, así como el resultado de introducirlo en el sistema.

- Promediado exponencial del histograma. Esta tarea surge como consecuencia de la variación de luminancia existente entre planos consecutivos, debido a la entrada y salida de sujetos en la escena. De modo que entre planos consecutivos los histogramas a ecualizar pueden sufrir cambios en mayor o menor medida, variando el contraste conseguido tras la ecualización de un plano al siguiente, y por tanto, la definición de contornos y realce de detalle.
  Se propone aplicar la ecualización, no sobre el histograma de la imagen actual, sino sobre un histograma promedio con respecto al histograma de la imagen de referencia, con el fin de suavizar las posibles variaciones.

Esta propuesta se desarrolla en el siguiente apartado.

- Explorar otras alternativas a la ecualización de histograma. Tal y como se ha podido observar en los resultados, los efectos de la ecualización en ciertas secuencias no son del todo efectivos. En particular, secuencias con un alto contenido de grises muy claros, tienden a oscurecer demasiado; un ejemplo muy representativo es "*Ice*". Y, por el contrario, en secuencias oscuras o con bandas negras, tras la ecualización, las regiones de interés son aclaradas demasiado. Estas observaciones dan lugar al planteamiento de técnicas alternativas que conformen la etapa de pre-procesado.

5.2.3.5 Mejora de la etapa de pre-procesado

En relación a las posibles mejoras de la etapa de pre-procesado propuestas en el apartado anterior, este apartado se centra en la implementación de una de ellas. En concreto, se decanta por la aplicación de un promediado exponencial en el histograma a ecualizar. Con la incorporación de esta herramienta se busca suavizar las variaciones de histogramas de planos consecutivos, con el fin de evitar en la medida de lo posible las consecuentes variaciones de contraste y detalle, antes de la ecualización.

La variación de contraste y detalle en los planos puede repercutir en el cálculo del mapa de vectores característicos y como consecuencia, en el mapa binario de clasificación. Si se consiguen eliminar total o parcialmente esas diferencias, se podrá aprovechar correctamente el detalle que realza la ecualización para generar una representación del movimiento más fidedigna.

Por tanto, para tratar este problema se recurre a una transformación del histograma de cada plano, en la que interviene el histograma (original) del plano anterior. El promediado exponencial utilizado responde a la siguiente expresión.

$$H'_k(i) = \alpha \cdot H_k(i) + (1 - \alpha) \cdot H'_{k-1}(i) \qquad (22)$$

Donde $H_k(i)$ es el histograma característico del plano actual, $H'_{k-1}(i)$ el histograma promediado correspondiente al plano anterior, y α el factor de olvido del promediado exponencial. Este coeficiente determina la importancia que tiene

el histograma del plano actual con respecto al del plano anterior y viceversa. Para establecer un valor predeterminado a este parámetro es necesario probar un rango adecuado de valores (, y) y analizar el efecto que producen; además, las secuencias utilizadas en este estudio deben ser aquellas que presenten diferencias de luminancia destacables por la entrada o salida de objetos en la escena. Por tanto, se recurre, por un lado, a los vídeos "*Pedestrian*" y "*Ice*" (resolución CIF) en los que existen idas y venidas de personas en el plano, y, por otro lado, "*África*" (SD), que además de tener mayor resolución, presenta un grado alto de detalle en el fondo.

Tras analizar los vídeos clasificados, generados a partir de los resultados de esta pequeña prueba, apenas se observan cambios en los mapas binarios con respecto a los obtenidos sin el promediado exponencial del histograma. La uniformidad y la coherencia temporal de la clasificación son similares en ambas versiones, por tanto, no aporta ninguna mejora en estos casos, aunque tampoco introduce ningún error. Como consecuencia de los resultados, se buscan en todos los vídeos de trabajo disponibles secuencias en las que entre en escena un objeto de tamaño grande, que ocupe casi todo el plano; esta situación sería la más adecuada para comprobar la necesidad del promediado, pues los histogramas deben variar considerablemente. Las secuencias encontradas fueron "*LOTR*" (CIF) y "*Bohemia2*" (SD), en las que la variación de contraste de un plano a otro mientras un objeto entra en escena es destacable. Por desgracia, los resultados obtenidos tampoco son muy significativos, pues los mapas de clasificación obtenidos con y sin el promediado del histograma son casi idénticos, y observando con detalle la secuencia ecualizada con el suavizado se puede apreciar la similitud de los planos en cuanto a la variación de contraste entre planos consecutivos.

En definitiva, el promediado del histograma no supone una mejora significativa en la clasificación, aunque es cierto que tampoco la empeora. En conjunto, la etapa de pre-procesado podría obviarse en el sistema por no proporcionar una mejora genérica para todos los tipos de secuencias con los que se trabaja, llegando incluso a empeorar la clasificación de alguno de ellos.

5.3 Mejora del clasificador binario de movimiento λ_1

5.3.1 Introducción

Uno de los problemas principales encontrados a lo largo del desarrollo del sistema de enmascaramiento por movimiento es la inconsistencia temporal en el mapa binario de clasificación, que puede tener consecuencias negativas sobre la calidad subjetiva del resultado final, haciendo que las variaciones de distorsión introducida sean perceptibles. Con el procesado morfológico la consistencia temporal se ve reforzada, consiguiendo mapas binarios más consistentes, tal y como puede observarse en las figuras adjuntas del apartado 5.1. Sin embargo, la coherencia temporal necesaria no es alcanzada totalmente, debido a que el clasificador de movimiento λ_1 consiste en un umbral fijo para todas las secuencias; y, la decisión no es muy consistente ante planos que presentan un mapa de vectores de valor en torno al umbral, a causa, por ejemplo, de una variación, por muy leve que sea, en el seguimiento del objeto de interés (en muchas ocasiones el cámara no sigue al objeto de interés con una velocidad de seguimiento totalmente constante); esto puede dar lugar a que MBs cosituados tengan una salida del clasificador distinta, cuando, en realidad, presentan un movimiento uniforme en dicha región.

Por otro lado, en el apartado 5.2, a partir de la introducción de la etapa de pre-procesado, se consigue una mejora en la clasificación en ciertas secuencias, pues como se concluyó, esta etapa no generaliza correctamente, de ahí las mejoras propuestas al final del apartado de la misma. Una de las soluciones planteadas ante este problema es dotar a la clasificación de movimiento de información espacio-temporal, es decir, que en la decisión también intervengan la clasificación realizada en MBs vecinos pertenecientes al mismo plano y el MB cosituado en la imagen anterior. De esta manera, la clasificación dejaría de ser dura, pasando a disponer de un rango de umbrales posibles para solventar los errores producidos ante variaciones de movimiento en torno al umbral λ_1. La solución propuesta se detalla en el siguiente apartado.

5.3.2 Descripción del algoritmo

La primera consideración temporal incorporada al clasificador λ_1 se localiza en la entrada al mismo. Recuérdese que en la primera versión del clasificador la entrada era simplemente el módulo del vector de movimiento asociado a cada macrobloque (|VM(i,j)|), pero la entrada del nuevo clasificador consiste en el promedio entre el módulo del vector del MB actual, denotado por $|VM(i,j)|_k$, y su cosituado, $|VM(i,j)|_{k-1}$, donde *k* indica el número de plano. Con este promedio se pretenden suavizar las posibles variaciones entre vectores de movimiento cosituados, producidas por las diferencias de velocidad del movimiento de cámara; de esta manera se busca reducir el efecto de estas situaciones sobre el mapa de clasificación. En definitiva, la expresión característica del promedio se presenta a continuación, y, como se puede observar, el factor de promediado que determina el peso de cada elemento se denomina β.

$$\left|VM(i,j)\right|_k = \beta \cdot \left|VM(i,j)\right|_k + (1-\beta) \cdot \left|VM(i,j)\right|_{k-1} \qquad (23)$$

Por tanto, el promedio del módulo del vector se considera la nueva entrada al clasificador. Éste cuenta con dos umbrales prefijados: uno superior, *UmbSup*, y otro inferior, *UmbInf*, que se comprueban al inicio de la clasificación; de modo que, si $|VM(i,j)|_k$ es superior o igual al umbral *UmbSup*, el MB correspondiente se clasifica como con mucho movimiento (a 1), y, si, por el contrario, es inferior o igual a *UmbInf*, se considera de poco movimiento (a 0).

Por otro lado, para aquellos MBs cuyos módulos se sitúen en el intervalo *(UmbInf, UmbSup)* existe una comprobación adicional. Al no cumplir ninguna de las condiciones anteriores, se puede deducir que se trata de vectores conflictivos cuya clasificación más adecuada puede resultar confusa, por lo que se recurre a la información disponible de los MB vecinos clasificados, así como a la clasificación del MB cosituado, generando una función de coste detallada en (24).

$$cost(i,j)_k = [\,] \cdot \lambda_1(i,j)_{k-1} + [\,] \frac{1}{N_{vecinos}} \sum_a \sum_b \lambda_1(i-a, j-b)_k \qquad (24)$$

El primer término de la función de coste se corresponde con la clasificación del MB cosituado. El segundo término indica la clasificación promedio de los vecinos ya clasificados; su valor se sitúa entre 0 y 1. Se puede observar en esta expresión que la clasificación del MB cosituado tiene un peso importante en la decisión de clasificación del MB actual, de un [ ], para conseguir una mayor

coherencia temporal; la responsabilidad de decisión restante se reparte entre cada uno de los MBs vecinos previamente clasificados.

Por último, el coste asociado a cada MB se compara con otro umbral, *UmbCost*. Para resumir cada uno de las condiciones que comprueba el clasificador se recurre al diagrama de la Figura 5.46, para facilitar la comprensión del funcionamiento del nuevo clasificador de movimiento.

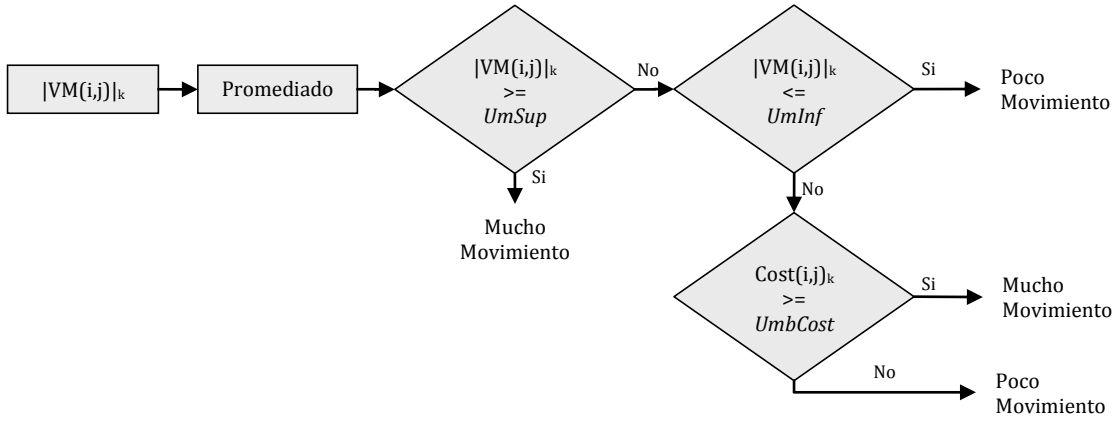


Figura 5.46. Diagrama de bloques del nuevo clasificador de movimiento, basado en información espacio-temporal.

Tras ofrecer una explicación del algoritmo característico del nuevo clasificador, es necesario configurar los parámetros nuevos que aporta, como los umbrales prefijados o el factor de promediado. Estas variables son de vital importancia en la consecución de un mapa binario de clasificación eficiente, pues, en función de su valor, restringen la decisión en menor o mayor medida y determinan el peso de la información temporal de la que se dispone. Todo esto se recoge en el siguiente apartado.

5.3.3 Configuración del algoritmo

Para comprobar la eficacia del nuevo clasificador basado en información espacio-temporal es necesario realizar un conjunto de pruebas que, en primer lugar, determinen los valores a asignar a los parámetros configurables que conforman el algoritmo. La versión del sistema utilizada es la correspondiente al algoritmo EMJ junto con la etapa de post-procesado; la etapa de procesado previa se encuentra inhabilitada puesto que requiere mejoras. Las secuencias a las que se recurre en la evaluación del clasificador son las utilizadas numerosas veces en pruebas anteriores, de modo que se dispone de variedad de resolución así como de

características de movimiento en la escena. A continuación, se señalan los parámetros configurables, y se añaden las imágenes representativas pertinentes.

- Factor de promediado, β

Los valores de prueba utilizados han sido [redacted], [redacted] y [redacted], asignando un peso cada vez mayor al módulo del vector de movimiento del MB en estudio. Los mapas binarios obtenidos en cada caso han sido muy similares, a pesar de la diferencia de peso aplicada en cada caso: para β=[redacted] el vector de movimiento del plano anterior adquiere gran importancia en el promediado, mientras con β=[redacted] se obtiene el caso contrario. Por tanto, por las similitudes entre las salidas del clasificador, se ha decidido seleccionar un valor de compromiso de [redacted].

- Umbrales *UmbSup* y *UmbInf*

Tras el promediado, el siguiente paso en el clasificador consiste en comprobar si se superan o no estos dos umbrales. La asignación de valores para estas dos variables no requiere la realización de pruebas con diferentes configuraciones, sino que se basa en los resultados obtenidos en las pruebas relacionadas con el apartado 4.2. Por tanto, los umbrales superior e inferior quedan prefijados según indica la Tabla 5.2 para cada uno de los tamaños de secuencia con los que se han trabajado.

	QCIF	**CIF**	**SD**
UmbSup	[redacted]	[redacted]	[redacted]
UmbInf	[redacted]	[redacted]	[redacted]

Tabla 5.1. Asignación valores *UmbSup* y *UmbInf*

Los umbrales se han establecido a dichos valores puesto que a lo largo de las pruebas realizadas en versiones anteriores del algoritmo el umbral del clasificador simple se ha mantenido entre estos pares de valores para cada caso en la mayoría de los vídeos de prueba, pues los módulos varían alrededor del valor intermedio.

- Umbral de la función de coste *UmbCost*

Por último, para fijar el umbral de coste se probaron dos valores: [redacted] y [redacted]. El primero resultó muy restrictivo porque para conseguir que el MB sea clasificado a 1, se debe cumplir que el MB cosituado esté clasificado a 1 y que la mayoría de los

MBs vecinos hayan sido clasificados a 1 también. Por ello, los mapas binarios generados eran erróneos. Sin embargo, con *UmbCost* igual a [redacted] la clasificación de los vecinos no es tan relevante, sino que independientemente de la misma, el MB cosituado marca el estado de la clasificación final, pues su peso en la función de coste (24) es importante; en este caso, la información de los vecinos es determinante sólo cuando el MB cosituado está clasificado a 0. Por tanto, el valor definitivo para *UmbCost* es [redacted], pues ofrece unos resultados más convincentes desde el punto de vista subjetivo.

Una vez seleccionados los valores más adecuados para cada parámetro configurable, se lanzaron unas pruebas de codificación con todas las secuencias utilizadas de costumbre con el objetivo de realizar una comparativa de los mapas binarios entre versiones del algoritmo desde el punto de vista subjetivo. Una observación importante que realizar consiste en la mejora en la consistencia temporal introducida por el nuevo clasificador, perceptible si se visualizan las secuencias clasificadas. Por otro lado, en ciertos vídeos se consigue delimitar mejor las regiones de interés; las siguientes figuras se corresponden con ejemplos demostrativos.

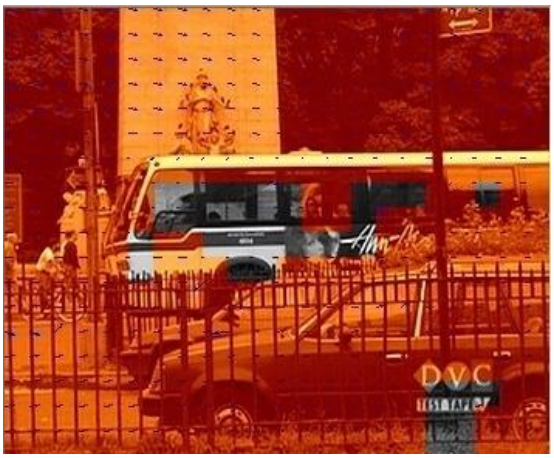


Figura 5.47. Salida clasificador λ_1 = [redacted] con post-procesado. Secuencia "Bus" (352x288).

Figura 5.48. Salida clasificador λ_1 mejorado con post-procesado. Secuencia "Bus" (352x288).

Figura 5.49. Salida clasificador $\lambda_1 =$ [redacted] con post-procesado. Secuencia "Stefan" (352x288).

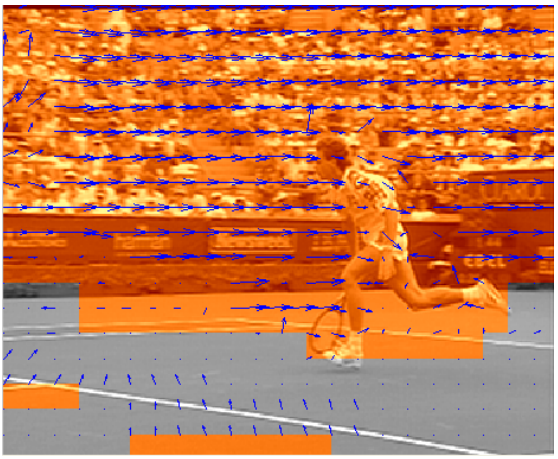


Figura 5.50. Salida clasificador λ_1 mejorado con post-procesado. Secuencia "Stefan" (352x288).

5.3.4 Conclusiones

A partir de los resultados de la codificación realizada en el apartado anterior, se puede concluir que la mejora introducida en el clasificador consigue una consistencia temporal más eficaz en los vídeos clasificados generados con respecto al clasificador simple (véase apartado 4.2); así, si se visualizan las secuencias correspondientes a esta mejora, por ejemplo, "bus_hys_morphoON.avi" o "pedestrian_hys_morphoON.avi", que están disponibles en el DVD adjunto, se observa como la variación en la clasificación de un plano al siguiente no es tan abrupta como antes, sino que los cambios se realizan de forma más gradual. Adicionalmente, delimita mejor las regiones de interés, tal y como puede observarse en las Figuras 5.47 a la 5.50; también, si se recurre al vídeo "bohemia_hys_morphoON.avi" se puede apreciar la mejora en la delimitación de las regiones de interés. Por todo ello, se espera que la distorsión introducida en zonas de mucho movimiento no sea perceptible, ya que esta distorsión está mejor localizada.

Capítulo 6

Estudio de la viabilidad del sistema mediante pruebas subjetivas

6.1 Introducción

En el capítulo 3 se detallan las partes de las que consta el sistema completo de enmascaramiento por movimiento. Los dos bloques principales que componen el sistema son, por un lado, el algoritmo EMJ y el primer clasificador λ_1, y por otro, el algoritmo de estimación de movimiento de cámara junto con el clasificador λ_2. En este trabajo aún no se cuenta con el segundo bloque del sistema, aunque, antes de proceder a su implementación y evaluación, es necesario realizar un estudio subjetivo de lo que se tiene hasta el momento, que proporcione indicios sobre la utilidad o no de dicho sistema. Hay que comprobar que el algoritmo ya implementado va a suponer una mejora en términos de reducción de tasa y, en caso afirmativo, hay que plantearse la necesidad de detectar regiones de interés para introducir grados de distorsión.

Por tanto, son necesarias pruebas subjetivas que determinen la viabilidad del sistema. En ellas, si se observa una reducción de tasa notable mientras la calidad subjetiva se mantiene, además de demostrar la validez del algoritmo, hay que probar la necesidad del segundo clasificador. Para conocer si el segundo clasificador es relevante o no, en este estudio se recurre a la evaluación subjetiva de dos grupos de vídeos, denominados Categoría 1 y Categoría 2 (véase descripción en apartado 6.2.1) codificadas con distorsiones similares. Si los dos grupos de vídeos se ven igual, esto indica que la cantidad de movimiento es lo único de lo que depende el enmascaramiento, y, por lo tanto, el clasificador λ_2 es irrelevante; esto se debe a que en ciertas secuencias, el mapa binario generado por el primer clasificador ya refleja una segmentación de las regiones de interés.

En este capítulo se describe la metodología de validación que se ha utilizado en dos pruebas subjetivas realizadas, en Julio de 2009 y Enero de 2010. El método de evaluación seleccionado lo propone [33], y consiste en una modificación de la medida VSSIM, junto con una variante de la MOS. A continuación, se detallan algunas decisiones tomadas sobre el diseño de las pruebas, en cuanto al conjunto de vídeos, la cantidad de distorsión, la selección de umbrales de clasificación…; toda la metodología y decisiones previas son comunes a ambas pruebas realizadas. También, se incluyen gráficos y tablas como presentación de los resultados correspondientes a cada una de las pruebas llevadas a cabo.

6.2 Decisiones previas del diseño de las pruebas

6.2.1 Selección de grupos de secuencias

Como se ha citado en el apartado anterior, se recurre a dos tipos de vídeos para realizar las pruebas. Estos conjuntos de secuencias surgen por la naturaleza del movimiento que presentan y como consecuencia, también, de los mapas binarios obtenidos en pruebas realizadas previamente.

La característica principal del primer conjunto de vídeos (Categoría 1) es que sólo los MBs con poco movimiento o nulo pertenecen a la zona de interés, es decir, la cámara sigue al objeto. Por su parte, el segundo grupo (Categoría 2) se

caracteriza por que las zonas de interés son aquellas que presentan un movimiento más elevado (no existe movimiento de cámara). Ejemplos representativos de ambos grupos pueden ser las secuencias "*Bus*" (Categoría 1) y "*Football*" (Categoría 2); en ellas, se puede observar que el clasificador λ_1 detecta regiones de interés, por tanto, el primer clasificador es suficiente, en principio, para evaluar la viabilidad del sistema completo.

Además, con estos dos grupos de vídeos es posible demostrar la validez del razonamiento teórico presentado en el capítulo 3 acerca de la cantidad de distorsión a introducir, pues, en caso de secuencias como "*Football*", a las zonas con movimiento rápido se les debería introducir menos distorsión que a las áreas que presentan movimiento en secuencias como "*Bus*", pues el ojo presta atención a los jugadores y no al fondo; por el contrario, en "*Bus*" el interés se centra en los bloques con movimiento nulo.

En definitiva, una vez implementado el algoritmo de clasificación de movimiento (capítulos 4 y 5) en el codificador H.264/AVC, la siguiente tarea que se debe llevar a cabo es la evaluación del sistema mediante pruebas subjetivas empleando los siguientes conjuntos de vídeos:

- **CATEGORÍA 1:** La cámara se mueve para seguir al objeto de interés.
  a. "Bus". Planos: 0-149
  b. "Bohemia". Planos: 65-201
  c. "Stefan". Planos: 172-299
  d. "Airshow3". Planos: 1439 – 1504

- **CATEGORÍA 2:** El movimiento de cámara es nulo y el objeto de interés se mueve.

  a. "Football". Planos: 0-124
  b. "Airshow1". Planos: 1092-1234
  c. "StarWars". Planos: 370-545
  d. "Pedestrian". Planos: 0 – 100

6.2.2 Selección del tipo y cantidad de distorsión a aplicar en las zonas enmascarables

Se espera que las pruebas subjetivas ayuden a elegir el tipo de distorsión a aplicar en regiones (o macrobloques) con elevado movimiento, que pueden ser

tratadas con peor calidad dado que el ojo apenas presta atención a dichas zonas. Existen dos tipos de distorsión posibles:

- **Distorsión ΔQP.** Aumenta el valor de la QP.
- **Distorsión IZZ *(Increased Zero Zone*).** Después de cuantificar los coeficientes de la DCT, a los que tengan un valor inferior a un umbral se les asigna el valor igual cero.

Tanto el valor de ΔQP como el del umbral IZZ han sido escogidos de forma personalizada para cada secuencia de prueba.

En la primera de las pruebas realizadas, la de Julio 2009, se descartó el uso del IZZ por los resultados aportados, por ello, la segunda tanda no contó con otra manera de introducir distorsión que el incremento de QP, aunque sí con 3 valores distintos de QP para estudiar el comportamiento de la opinión subjetiva con ΔQP.

6.2.3 Selección de las tasas objetivo

Otra de las variables presentes en la batería de pruebas subjetivas es qué tasa objetivo emplear para codificar las secuencias de test. Efectivamente, interesa evaluar la calidad de las secuencias para varias tasas objetivo, pero al existir cierta limitación en cuanto a la cantidad de secuencias a evaluar, se han escogido dos rangos: **tasa alta y tasa baja**.

También ha de tenerse en cuenta que la curva de calidad (ya sea opinión subjetiva o PSNR) se estanca para valores altos de tasa, de manera que las diferencias que puedan producirse en esa región entre distintas codificaciones serán muy marginales. Es por ello que el experimento subjetivo se ha llevado a cabo fundamentalmente para calidades bajas, entendiendo como tal un valor personalizado de tasa para cada secuencia (no es lo mismo la calidad de "*Airshow1*" a 256 kbps que la de "*Football*", una secuencia mucho más complicada), y sólo se han lanzado pruebas a calidades altas para un par de secuencias con el fin de ver si se cumple la hipótesis.

6.2.4 Selección umbral λ1

El valor del umbral λ_1 escogido en la primera de las pruebas no es uno genérico, sino que se buscó un valor adecuado para cada secuencia que permitiera

mejorar la consistencia temporal y, de este modo, obtener una salida más fiable del clasificador.

Sin embargo, la evolución del algoritmo permitió emplear en la segunda prueba subjetiva una técnica de selección de umbral automática utilizada por todas las secuencias que ofrece un mapa de decisión más coherente (véase el apartado 5.3).

6.2.5 Descripción de la metodología de las pruebas

Para realizar las pruebas se han consultado las recomendaciones correspondientes de la ITU [29, 30] que, aunque tratan de detallar exhaustivamente las condiciones en las que han de llevarse a cabo los experimentos, comprenden numerosas aplicaciones objetivo y han tenido que ser adaptadas a esta aplicación particular, obteniéndose las siguientes condiciones de trabajo:

- De entre los experimentos que se proponen en la recomendación para la evaluación de la opinión subjetiva se ha seleccionado el Índice por categorías de degradación (DCR), en el que se somete al sujeto a un doble estímulo: por un lado la secuencia de referencia (convenientemente identificada) y por otro la secuencia de prueba (que en este caso consistirá en la misma secuencia de vídeo codificada y reconstruida con distintas versiones del codificador). En este estudio, el DCR se llevará a cabo en modo *Simultaneous Presentation* (SP). El motivo principal para la elección de este tipo de experimento es la reducción del tiempo del mismo al presentar simultáneamente la referencia y la secuencia de test. Asimismo, se eliminan de la evaluación del espectador aspectos subjetivos al tener que evaluar simplemente la degradación entre dos estímulos.
- La elección de este tipo de experimento tiene como consecuencia el uso de un tamaño suficientemente reducido como para mostrar simultáneamente las dos secuencias en el monitor de test, así como la sincronización perfecta del inicio de ambas y el uso de un monitor suficientemente grande. Por tanto, se emplearon monitores planos de 19 pulgadas.

- La distancia del observador al monitor de test ha de ser de alrededor de 8 veces la altura de las imágenes, aunque la recomendación es bastante laxa al respecto, y el fondo de la pantalla ha de ser un gris al 50%, puesto que está demostrado que dependiendo de si el fondo es blanco o negro esto puede afectar a la percepción de determinados valores de brillo en las secuencias de test. Asimismo, se reduce la iluminación de la sala de test.

- Se incluye una prueba de agudeza visual (test de *Snellen*) y un test de *Ishiara* para detectar el daltonismo. Se descartarán aquellos individuos con un determinado número de fallos en alguna de las pruebas.

- El tiempo total del experimento es de unos 10-12 minutos para cada individuo, pudiendo realizarse simultáneamente hasta 8 pruebas en los equipos del laboratorio. La recomendación indica que se necesita un mínimo de 15 individuos para sacar conclusiones válidas y recomienda emplear un máximo de 40, puesto que según la recomendación los resultados no mejoran con un número mayor que éste.

- Se han empleado según la recomendación algunas secuencias de entrenamiento, para que el espectador calibre su opinión, y que serán mostradas en primer lugar. Además, se han realizado dos repeticiones para comprobar la fiabilidad de la opinión de cada sujeto. Si en dichas repeticiones se observa una diferencia de opinión muy significativa, hay motivos para descartarlo.

- Una vez eliminados aquellos sujetos que quedan descartados por su falta de agudeza visual o por los problemas para diferenciar los colores, se procede a eliminar a aquellos que han puntuado de manera muy distinta los experimentos repetidos y a aquellos que pudieran tener alguna experiencia previa en codificación de vídeo (que han sido convenientemente detectados mediante un formulario previo al experimento).

6.3 Experimento subjetivo número 1 Julio 2009

A continuación se incluye una tabla resumen de las secuencias, tasas y opciones de codificación del primer experimento subjetivo, del que incluiremos el análisis a modo de comparativa.

SECUENCIA	*TASA*	*Versión*
Bohemia	Baja	NCP
Bohemia	Baja + 20%	NCP
Bohemia	Baja	IZZ
Bohemia	Baja	DQP
Bus	Baja	NCP
Bus	Baja	IZZ
Bus	Baja	DQP
Stefan	Baja	NCP
Stefan	Baja	IZZ
Stefan	Baja	DQP
Stefan	Alta	NCP
Stefan	Alta	IZZ
Stefan	Alta	DQP
Football	Baja	NCP
Football	Baja + 20%	NCP
Football	Baja	IZZ
Football	Baja	DQP
Airshow	Baja	NCP
Airshow	Baja	IZZ
Airshow	Baja	DQP
Airshow	Alta	NCP
Airshow	Alta + 20%	NCP
Airshow	Alta	IZZ
Airshow	Alta	DQP
StarWars	Baja	NCP
StarWars	Baja	IZZ
StarWars	Baja	DQP

Tabla 6.1. Tabla resumen Experimento Subjetivo 1.

6.3.1 Resultados a tasa baja

En la primera categoría de vídeos el objeto de interés está siendo seguido por la cámara, de manera que el movimiento elevado se observa en el fondo. La Figura 6.1 muestra los resultados obtenidos en opinión subjetiva, que también se resumen en la siguiente tabla junto con el intervalo de confianza al 95%:

	Bus	Bohemia	Stefan	Incr. Medio sobre NCP
NCP*	3.21 ± 0.44	1.73 ± 0.3	2.85 ± 0.47	0
deltaQP	3.04 ± 0.34	2.21 ± 0.34	3.62 ± 0.35	0.36
IZZ	3.01 ± 0.42	2.14 ± 0.27	3.45 ± 0.41	0.27

Tabla 6.2. Comparativa de resultados para secuencias Categoría 1 (tasa baja).

*No Consideraciones Perceptuales

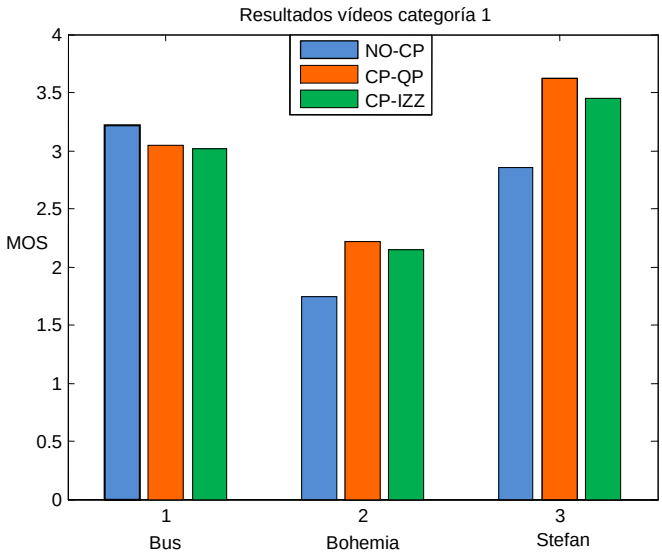


Figura 6.1. Resultados para las secuencias de vídeo Categoría 1: "Bus", "Bohemia" y "Stefan".

Como se puede observar, "*Bus*" es el único caso en que la aplicación de las consideraciones perceptuales no provoca un aumento significativo de la calidad observada, e incluso en este caso parece que la degradación que se obtiene no es muy significativa. Centrando la atención en el incremento medio de MOS que supone el uso de las técnicas, para la categoría 1 parece conveniente un tratamiento de este tipo.

De las dos técnicas de introducción de distorsión empleadas, la que consigue un resultado más llamativo para esta categoría de secuencias es el uso de la delta QP en lugar de la técnica IZZ, aunque los resultados no son muy concluyentes en este sentido.

Por otro lado, la segunda categoría de vídeos la componen secuencias en que el movimiento elevado se produce en el objeto de interés. Sus resultados se muestran en la Figura 6.2, así como en la siguiente tabla junto con el intervalo de confianza al 95%:

	Airshow	Star Wars	Football	Incr. Medio sobre NCP
NCP	3.02 ± 0.33	2.92 ± 0.3	3.38 ± 0.31	0
deltaQP	2.54 ± 0.46	2.57 ± 0.3	3.80 ± 0.45	-0.13
IZZ	3.02 ± 0.43	2.78 ± 0.43	3.55 ± 0.55	0.01

Tabla 6.3. Comparativa de resultados para secuencias Categoría 2 (tasa baja).

*No Consideraciones Perceptuales

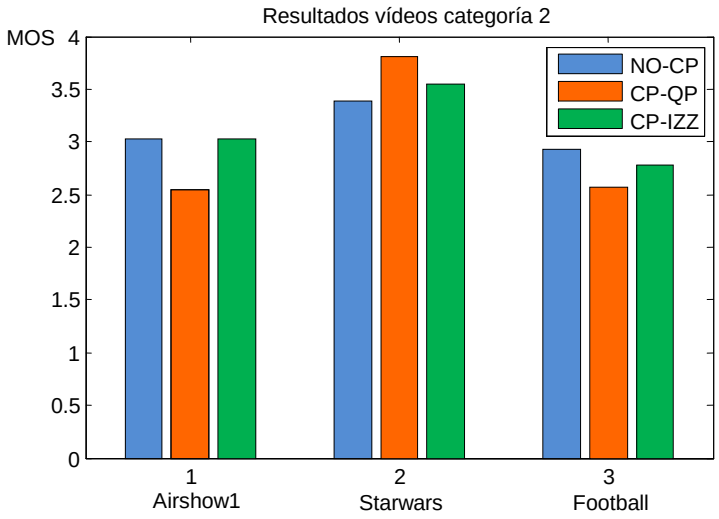


Figura 6.2. Resultados para las secuencias de vídeo de la categoría 2: "Airshow1", "Football" y "StarWars".

En este caso, para la categoría 2, los resultados no son tan buenos. Se observa cómo la opinión subjetiva se deteriora con el uso de las técnicas de enmascaramiento por movimiento, salvo quizás en la secuencia "*Football*". El resultado coincide con lo esperado, ya que en este caso, sin movimiento de cámara, las regiones que se mueven mucho serán objeto de una mayor atención y, por tanto, la introducción en ellas de una distorsión será más notable para el

espectador. Este resultado hace patente que el algoritmo ha de combinarse con otro basado en determinar el objeto de interés (aunque sea también según consideraciones de movimiento de cámara).

6.3.2 Resultados a tasa alta

Un vídeo de cada categoría ha sido codificado con tasa más alta, para verificar que realmente se cumple la hipótesis antes mencionada de que las diferencias no van a ser tan significativas. Se incluyen los resultados ya expuestos de tasa baja para comparar.

- Vídeo Categoría 1: "Stefan"

	NO-CP	CP-QP	CP-IZZ
MOS (tasa alta)	4.33 ± 0.33	4.80 ± 0.39	4.42 ± 0.49
MOS (tasa baja)	2.85 ± 0.47	3.62 ± 0.35	3.45 ± 0.41

Tabla 6.4. Comparativa MOS a tasa alta para la secuencia "Stefan".

- Vídeo Categoría 2: "Airshow1"

	NO-CP	CP-QP	CP-IZZ
MOS (tasa alta)	4.95 ± 0.43	4.71 ± 0.29	4.77 ± 0.46
MOS (tasa baja)	3.02 ± 0.33	2.54 ± 0.46	3.02 ± 0.43

Tabla 6.5. Comparativa MOS a tasa alta para la secuencia "Airshow1".

6.3.3 Resultados para un aumento de tasa del 20%

El siguiente experimento pretendía ilustrar si la opinión subjetiva ante una secuencia codificada con consideraciones perceptuales puede asimilarse a la que tendría el espectador a la vista de la secuencia sin consideraciones perceptuales codificada con un cierto incremento de tasa (en este caso se ha probado un 20%). Esto se ha hecho tanto a tasa baja, como a tasa alta y con ambas categorías de vídeos. Los resultados son los siguientes:

- Vídeo Categoría 1: "Bohemia". Tasa baja:

	NO-CP	NO-CP + 20%	CP-QP	CP-IZZ
MOS	1.73	2.14	2.21	2.14

Tabla 6.6. Resultados experimento Aumento de Tasa del 20%. Categoría 1, Tasa baja.

- Vídeo Categoría 2: "Football". Tasa baja:

	NO-CP	NO-CP + 20%	CP-QP	CP-IZZ
MOS	3.38	4.10	3.81	3.55

Tabla 6.7. Resultados experimento Aumento de Tasa del 20%. Categoría 2, Tasa baja.

- Vídeo Categoría 1: "Stefan". Tasa alta:

	NO-CP	NO-CP + 20%	CP-QP	CP-IZZ
MOS	4.33	4.45	4.81	4.43

Tabla 6.8. Resultados experimento Aumento de Tasa del 20%. Categoría 1, Tasa alta.

- Vídeo Categoría 2: "Airshow1". Tasa alta:

	NO-CP	NO-CP + 20%	CP-QP	CP-IZZ
MOS	4.95*	4.79*	4.71	4.77

Tabla 6.9. Resultados experimento Aumento de Tasa del 20%. Categoría 2, Tasa alta.

*En ciertos vídeos la MOS no parece fiable, porque con un aumento de tasa del 20% la opinión subjetiva empeora.

De nuevo en la categoría 1 se obtiene un resultado positivo, pues los vídeos codificados sin consideraciones perceptuales pero con un 20% de incremento de tasa tienen una calidad comparable a la de los que sí usan consideraciones perceptuales con IZZ, e incluso pierden con respecto a los que usan consideraciones perceptuales con incremento de QP. En la categoría 2, sin embargo, no se puede sacar la misma conclusión.

6.3.4 Resultados del índice MOVIE

Con el fin de determinar si es posible aproximar la opinión subjetiva emitida por los espectadores con un indicador obtenido de manera objetiva, se ha probado una de las más modernas medidas objetivas de calidad subjetiva, el índice MOVIE (ver [31]), del que se proporcionan unas pautas generales en el apartado 2.3.6.

Dicho índice es más bien una medida de distorsión, puesto que otorga una puntuación de 0 a las secuencias idénticas a la original y una puntuación de 1 a las más distorsionadas, pero nótese que lo que se encuentra representado en las gráficas es una versión modificada del MOVIE para que ofrezca resultados entre 0 y 1, donde 0 es la peor calidad y 1 la mejor, para poder comprarlo con la MOS.

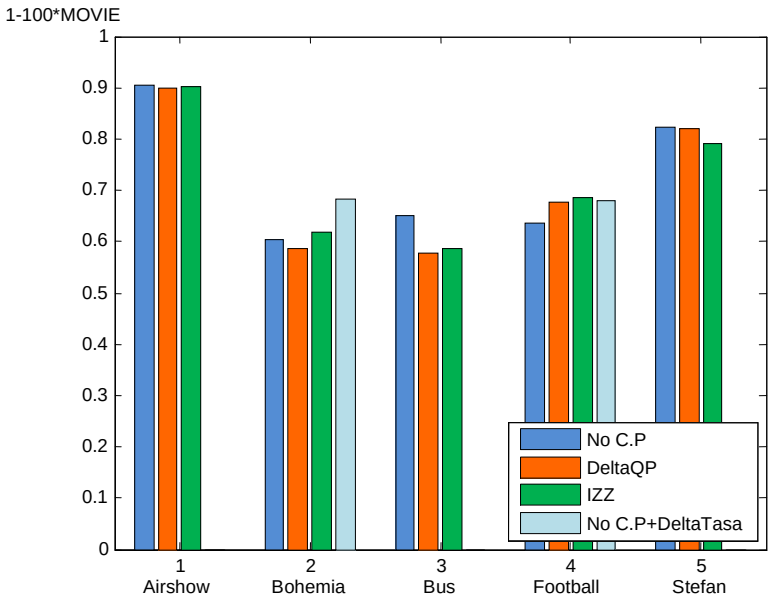


Figura 6.3. Medida de calidad basada en índice MOVIE para tasa baja.

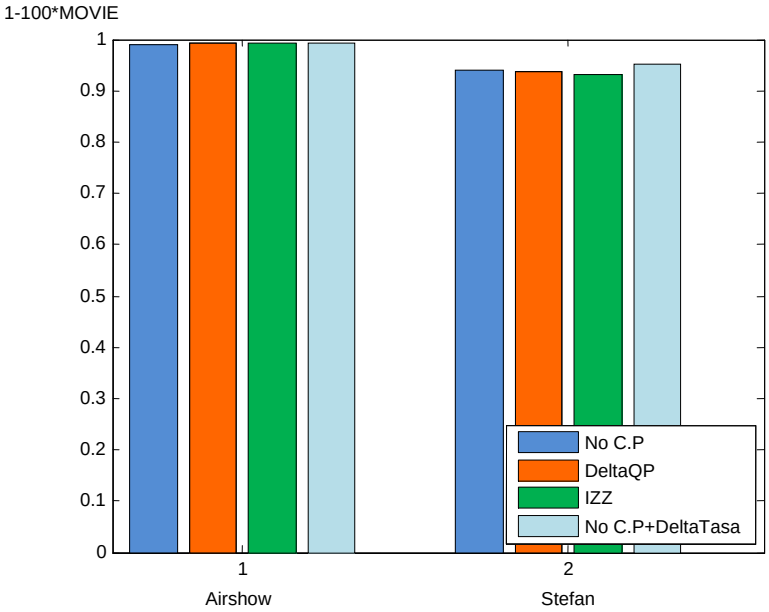


Figura 6.4. Medida de calidad basada en índice MOVIE para tasa alta

6.3.5 Análisis de los Resultados

A continuación se resumen los resultados obtenidos en el primer experimento subjetivo y que dieron lugar a la evolución del algoritmo:

- En los vídeos de categoría 1 la MOS mejora con consideraciones perceptuales (salvo en la secuencia "*Bus*"), mientras que los resultados son peores con las secuencias de categoría 2 (salvo en la secuencia "*Football*"). La caracterización de fondo-objeto.

- El método de introducción de distorsión mediante DeltaQP mejora al IZZ, al menos en los vídeos de Categoría 1 que son los que interesan a la vista de los resultados.

- La codificación de secuencias a tasa alta ha demostrado también que incluso en calidades más altas las diferencias en la MOS son apreciables y siguen la misma tendencia que para tasas bajas, al contrario de lo que inicialmente se pensaba sobre la apreciación de las diferencias únicamente a tasa baja.

- Por último, y como prueba tentativa para determinar el rango de ahorro de tasa del que se podría estar hablando para las consideraciones perceptuales, se observa en las pruebas de incremento del 20% en la tasa sin consideraciones perceptuales que para los vídeos de categoría 1 la MOS sigue siendo mejor en DeltaQP, tanto para tasa alta como para tasa baja. En los vídeos de categoría 2 no se produce tal incremento.

- Adicionalmente, se puede interpretar el comportamiento de la MOS como poco correlacionado con la opinión subjetiva. A la vista de las barras que representan (1 - 100 · MOVIE) sólo coincide plenamente con la MOS en el caso de "*Airshow*" (en este caso sólo da el mismo ganador) y parcialmente en "*Bus*" y "*Football*". En el resto de secuencias se comporta bastante mal e incluso totalmente opuesta a la opinión subjetiva.

6.4 Experimento subjetivo número 2 Enero 2010

Con el fin de evaluar la validez de la versión actual en ese momento del algoritmo de enmascaramiento por movimiento se llevó a cabo una batería de pruebas subjetivas para obtener la MOS de los vídeos de test. Concretamente se probaron varias configuraciones del algoritmo, que contienen (o no) los bloques de pre-procesado y post-procesado, varios incrementos de QP a aplicar en zonas de mucho movimiento ([redacted], [redacted] y [redacted]) y dos tasas objetivo (una baja y otra alta). La siguiente tabla resume el conjunto de vídeos de prueba escogido para la batería de pruebas subjetivas (se han obviado las secuencias sin consideraciones perceptuales).

SECUENCIA	PRE-PROC	POS-PROC	TASA	DELTA_QP
Bohemia (SD)	SIN	CON	Baja (400)	[redacted]
Bohemia (SD)	SIN	CON	Baja (400)	[redacted]
Bohemia (SD)	SIN	CON	Baja (400)	[redacted]
Bohemia (SD)	CON	CON	Baja (400)	[redacted]
Bohemia (SD)	SIN	SIN	Baja (400)	[redacted]
Bus (CIF)	SIN	CON	Baja (192)	[redacted]
Bus (CIF)	SIN	CON	Baja (192)	[redacted]
Bus (CIF)	SIN	CON	Baja (192)	[redacted]
Stefan (CIF)	SIN	CON	Baja (256)	[redacted]
Stefan (CIF)	SIN	CON	Baja (256)	[redacted]
Stefan (CIF)	SIN	CON	Baja (256)	[redacted]
Stefan (CIF)	SIN	CON	Alta (1024)	[redacted]
Stefan (CIF)	SIN	CON	Alta (1024)	[redacted]
Stefan (CIF)	SIN	CON	Alta (1024)	[redacted]
Football (CIF)	SIN	CON	Baja (512)	[redacted]
Football (CIF)	SIN	CON	Baja (512)	[redacted]
Football (CIF)	SIN	CON	Baja (512)	[redacted]
Football (CIF)	SIN	CON	Alta (1536)	[redacted]
Football (CIF)	SIN	CON	Alta (1536)	[redacted]
Football (CIF)	SIN	CON	Alta (1536)	[redacted]
Pedestrian(CIF)	SIN	CON	Baja (128)	[redacted]

Pedestrian(CIF)	SIN	CON	Baja (128)	
Pedestrian(CIF)	SIN	CON	Baja (128)	
Pedestrian(CIF)	SIN	SIN	Baja (128)	
Pedestrian(CIF)	CON	CON	Baja (128)	
Airshow2 (SD)	SIN	CON	Baja (300)	
Airshow2 (SD)	SIN	CON	Baja (300)	
Airshow2 (SD)	SIN	CON	Baja (300)	
Airshow2 (SD)	CON	CON	Baja (300)	

Tabla 6.10. Tabla resumen Experimento Subjetivo 2.

En las figuras y tablas que se incluyen a continuación se resumen los datos correspondientes a la opinión subjetiva media en las secuencias estudiadas y para todos los valores posibles de deltaQP. Estas gráficas se acompañan de las correspondientes al índice MOVIE a efectos comparativos. Se ha añadido a la opinión subjetiva MOS su intervalo de confianza al 95%.

6.4.1 Vídeos de categoría 1 a tasa baja

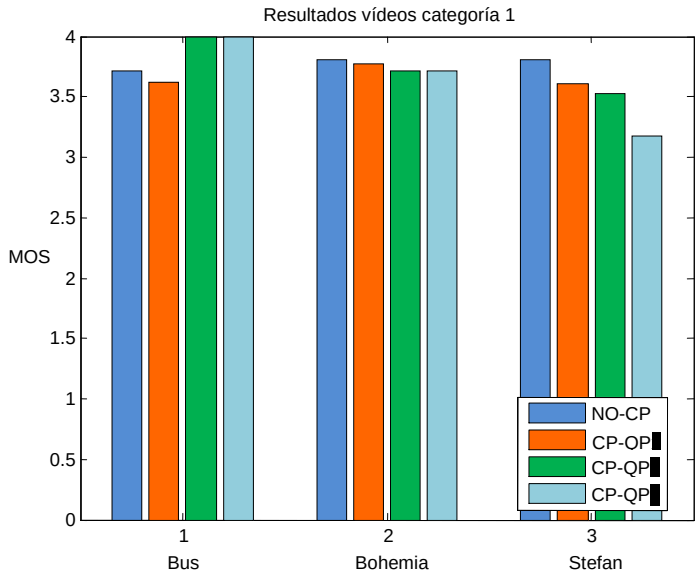


Figura 6.5. Medium Opinion Score (MOS) para secuencias de categoría 1 en tasa baja

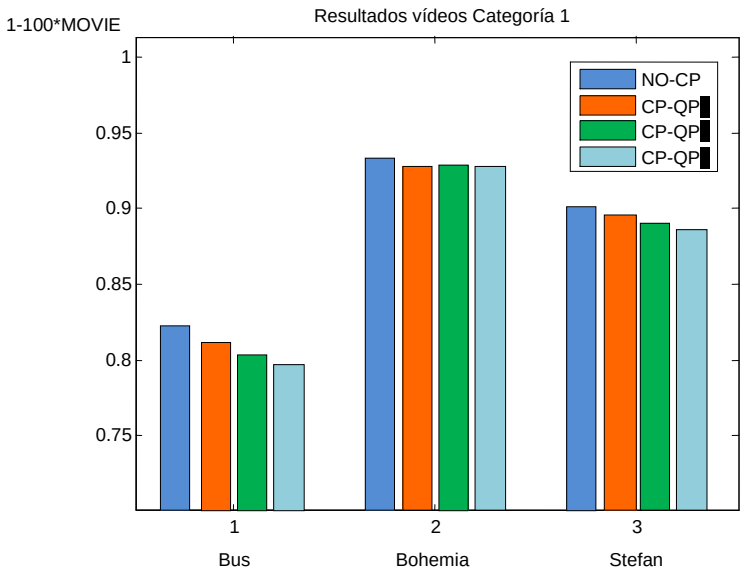


Figura 6.6. Medida de calidad basada en MOVIE para secuencias de categoría 1 en tasa baja

- Media MOS Categoría 1, tasa baja:

	MOS Sin CP	MOS DQP=[redacted]	MOS DQP=[redacted]	MOS DQP=[redacted]
Bus	3.7 ± 0.39	3.62 ± 0.30	4.00 ± 0.28	4.00 ± 0.28
Bohemia	3.81 ± 0.49	3.77 ± 0.33	3.71 ± 0.42	3.71 ± 0.23
Stefan	3.81 ± 0.34	3.60 ± 0.24	3.52 ± 0.23	3.18 ± 0.30
Media	**3.77**	**3.66**	**3.74**	**3.63**

Tabla 6.11. Media MOS de secuencias de Categoría 1, Tasa baja.

6.4.2 Vídeos de categoría 2 a tasa baja

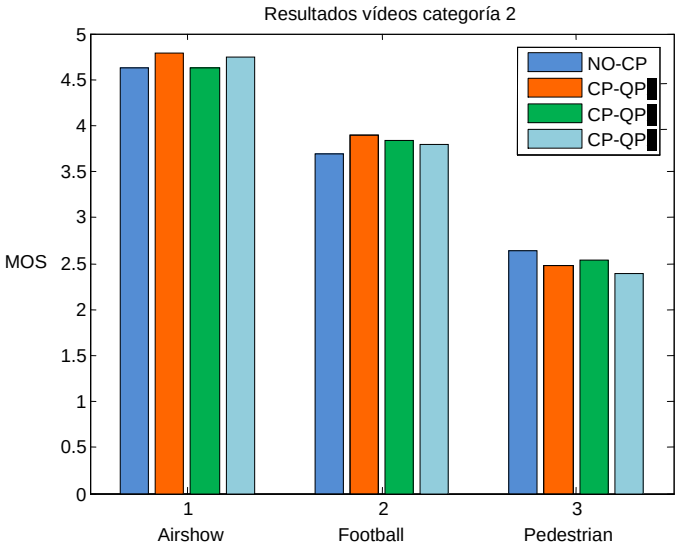


Figura 6.7. Medium Opinion Score (MOS) para secuencias de categoría 2 en tasa baja

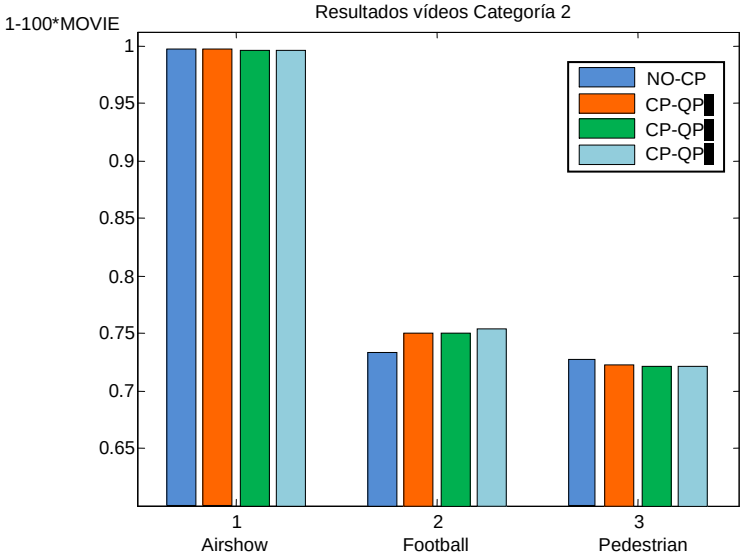


Figura 6.8. Medida de calidad basada en MOVIE para secuencias de categoría 2 en tasa baja

- Media MOS Categoría 2, tasa baja:

	MOS Sin CP	MOS DQP=[redacted]	MOS DQP=[redacted]	MOS DQP=[redacted]
Airshow	4.62 ± 0.30	4.79 ± 0.38	4.62 ± 0.26	4.75 ± 0.26
Football	3.69 ± 0.23	3.40 ± 0.35	3.83 ± 0.43	3.79 ± 0.24
Pedestrian	2.65 ± 0.27	2.48 ± 0.23	2.54 ± 0.32	2.40 ± 0.13
Media	**4.15**	**3.56**	**3.66**	**3.65**

Tabla 6.12. Media MOS de secuencias de Categoría 1, Tasa baja.

6.4.3 Vídeos a tasa alta (uno de cada categoría)

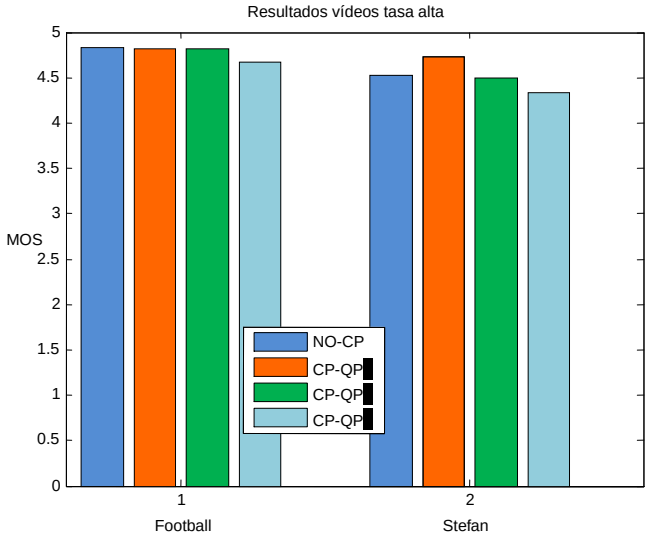


Figura 6.9. Medium Opinion Score (MOS) para tasa alta

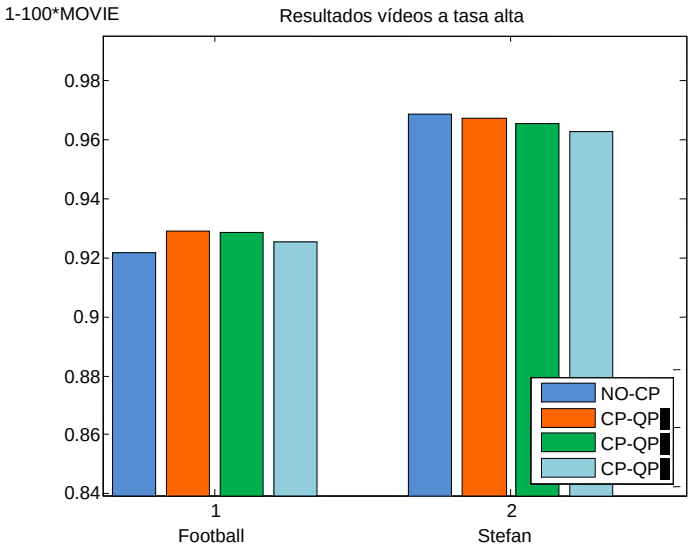


Figura 6.10. Medida de calidad basada en MOVIE para tasa alta

- Media MOS. Tasa alta:

	MOS Sin CP	MOS DQP=[redacted]	MOS DQP=[redacted]	MOS DQP=[redacted]
Football	4.83 ± 0.39	4.81 ± 0.29	4.81 ± 0.38	4.67 ± 0.24
Stefan	4.52 ± 0.24	4.73 ± 0.37	4.50 ± 0.22	4.33 ± 0.47

Tabla 6.13. Media MOS. Tasa alta.

6.4.4 Pre-procesado y post-procesado

Mientras que en la versión básica, cuyos resultados se han mostrado hasta ahora, el pre-procesado está desactivado y el post-procesado activado, se ha intentado establecer la importancia de dichos bloques combinándolos de maneras diferentes para dos secuencias particulares. Los resultados se resumen a continuación, junto con los correspondientes a no emplear consideraciones perceptuales y a la configuración básica de Pre-procesado NO y Post-procesado SI.

- Resultados Pre-Procesado NO, Post-Procesado NO y DeltaQP=[redacted]

	MOS Sin CP	MOS Pre NO Post SI DQP=[redacted]	MOS Pre NO Post NO DQP=[redacted]	1-100*MOVIE Sin CP	1-100*MOVIE Pre NO PostSI DQP=[redacted]	1-100*MOVIE Pre NO Post NO DQP=[redacted]
Pedestrian	2.65	2.54	2.35	0.7275	0.7212	0.7150
Bohemia	3.81	3.71	3.73	0.9330	0.9281	0.9282

Tabla 6.14. Comparativa 1 de resultados con diferentes combinaciones Pre y Post.

- Resultados Pre-Procesado SÍ, Post-Procesado SÍ y DeltaQP=[redacted]

	MOS Sin CP	MOS Pre NO Post SI DQP=[redacted]	MOS Pre SI Post SI DQP=[redacted]	1-100*MOVIE Sin CP	1-100*MOVIE Pre NO Post SI DQP=[redacted]	1-100*MOVIE Pre SI Post SI DQP=[redacted]
Airshow	4.62	4.62	4.79	0.9969	0.9965	0.9967
Bohemia	3.81	3.71	3.64	0.9330	0.9281	0.9245

Tabla 6.15. Comparativa 2 de resultados con diferentes combinaciones Pre y Post.

6.4.5 Análisis de los Resultados

A continuación se detallan algunas de las consecuencias de los resultados anteriores, que serán comentadas en más profundidad en las conclusiones.

- En los vídeos de categoría 1 los resultados son distintos de los de Julio. Se han obtenido reducciones en la MOS por efecto de las consideraciones perceptuales para "*Bohemia*" y "*Stefan*", y únicamente un aumento en "*Bus*". Las secuencias de categoría 2 consiguen buenos resultados en "*Airshow3*" y "*Football*" (en las pruebas de Julio "*Football*" también mejoraba cuando se aplicaban consideraciones perceptuales),

mientras que "*Pedestrian*" obtiene una MOS superior sin consideraciones perceptuales.

- En cuanto a la cuantía de la distorsión a introducir, se observa que los valores más adecuados de deltaQP están en torno a x[redacted], obteniendo las mayores mejoras y las menores pérdidas en la mayoría de los vídeos.

- Los resultados en tasa alta parecen desdecir a los de tasa baja. La codificación de "*Football*" con consideraciones perceptuales sale ahora perdiendo, mientras que la de "*Stefan*" ahora mejora al codificador básico, al contrario de lo que ocurría en calidades bajas. Sí se observa quizás que las diferencias son menos significativas que en el caso de calidades bajas.

- Otra de las opciones probadas consiste en eliminar del algoritmo la intervención del post-procesado, por ver en qué manera influye su uso. En los resultados obtenidos podemos observar cómo la secuencia "*Pedestrian*" se codifica mejor con post-procesado mientras "*Bohemia*" se habría codificado ligeramente mejor sin él (aunque las diferencias en la MOS y MOVIE son ínfimas). Esta pequeña diferencia podría deberse a que la etapa de post-procesado delimita algo peor los objetos de interés en este tipo de secuencias.

- Además, se ha probado la incorporación de un bloque de pre-procesado a las consideraciones perceptuales, que permitiría una mejora en la estimación de movimiento jerárquica en condiciones de bajo contraste. En este caso las cifras nos dicen que dicho bloque beneficiaría a la MOS de "*Airshow3*" mientras que perjudicaría a la de "*Bohemia*".

Capítulo 7

Conclusiones y trabajo futuro

Las conclusiones recogidas en este capítulo se extraen directamente del análisis de resultados de ambas pruebas subjetivas realizadas, encargadas de evaluar la viabilidad del sistema propuesto. En primer lugar, se listan las observaciones relativas a la prueba realizada en Julio 2009, en relación a la primera versión del algoritmo disponible, que no contaba aún con las etapas de pre y post-procesado, ni con la mejora del clasificador λ_1. Posteriormente, se añaden las ideas más destacadas sobre los resultados de la segunda prueba subjetiva. Y, para finalizar, se aporta una serie de posibles líneas futuras de trabajo con las que se podría continuar el estudio.

7.1 Conclusiones de las pruebas subjetivas de Julio

- La primera observación destacada que se extrae de las pruebas subjetivas es que aquellas secuencias en las que existe un seguimiento del objeto de interés por parte de la cámara y que tienen, por tanto, movimiento elevado

en el fondo, presentan una mejora significativa cuando se aplican consideraciones perceptuales en la codificación; por su parte, las secuencias pertenecientes a la otra categoría, en las que no existe movimiento de cámara y el movimiento se concentra en las regiones de interés, la distorsión introducida es muy notable y las consideraciones perceptuales no aportan ninguna mejora con respecto a la codificación original.

- En cuanto a la manera de introducir distorsión, se ha demostrado que el algoritmo de incremento de la zona de cero del cuantificador (IZZ) proporciona peores resultados que el incremento de QP, al contrario de lo que se esperaba, puesto que el incremento de QP puede generar efecto de bloques. Por lo tanto, el IZZ queda descartado, porque no es capaz de reducir significativamente el número de bits empleados en las zonas con enmascaramiento para luego usarlos en el resto de la secuencia.

7.2 Conclusiones de las pruebas subjetivas de Enero

- La validez de los resultados de esta última prueba subjetiva puede ser cuestionada, puesto que éstos difieren mucho con respecto a los resultados de la prueba subjetiva de Julio, a pesar de no haber modificado la filosofía del algoritmo, sino que se han añadido mejoras en la consistencia visual de salida. Si se observan los resultados, el algoritmo en lugar de mejorar, parece conseguir todo lo contrario. De hecho, conversando con algunos sujetos que se habían sometido a las pruebas comentaron que en secuencias correspondientes a la categoría 1 fijaban su atención en partes del plano situadas más allá de la región de interés que se había supuesto. Todas las irregularidades encontradas tras el análisis de resultados pueden haber sido causadas por diferentes razones que se exponen a continuación:

    - Una de las causas de estos resultados insatisfactorios puede ser que realmente el algoritmo haya empeorado. Si no se consigue determinar la región de interés de manera específica, es que el modelo simple que se ha asumido no es válido. Las supuestas mejoras añadidas al algoritmo han sido la modificación del clasificador y las etapas de pre y post-procesado. Dado que los

resultados sin estos cambios no son concluyentes, tampoco se puede determinar que estas nuevas etapas son la causa de la pérdida de MOS sufrida de la primera prueba a la última realizada. Aunque existe la posibilidad de que la causa fuera la etapa de post-procesado, debido a que las regiones de interés en los vídeos de categoría 1 no quedan bien delimitados; además, por su parte, el clasificador a pesar de ser algo más adaptativo no proporciona mapas de zonas enmascarables muy diferentes a los que devolvía antes.

- Otra posibilidad es que la técnica de introducción de la distorsión no sea la adecuada, pues, al existir el procesado de la nueva versión éste puede acentuar la degradación de la calidad, haciéndolo más perceptible que en las pruebas anteriores. Como ahora se dispone de un mapa binario más consistente en el tiempo, puede que la pérdida de detalle se acentúe por el uso de un escalón de cuantificación elevado; sin embargo, en la versión anterior del algoritmo, al no existir tanta coherencia temporal entre mapas consecutivos, se alternaban los distintos valores de QP de un plano al siguiente conservándose algo más de detalle por la existencia de una especie de "refresco".

- Además, es probable que el mapa binario que determina las zonas enmascarables no sea muy acertado, de modo que se esté introduciendo distorsión en zonas de interés subjetivo, haciendo necesaria la implementación del segundo clasificador que delimite las regiones de interés.

- Como los resultados de Julio y Enero no son consistentes, cabe poner en duda la fiabilidad de las pruebas subjetivas, y por tanto, las conclusiones basadas en ellas. Las observaciones que dan lugar a esta teoría son que el índice MOVIE mantiene cierta correlación con la MOS en las pruebas de Enero, pero en las de Julio no sucede lo mismo. Un aspecto de diseño que ha podido influir en estos resultados es, la cantidad de vídeos a visionar, que han aumentado con respecto a la prueba de Enero, aunque la duración de la prueba no aumentó demasiado. Esto provoca que el individuo al ver en

numerosas ocasiones la misma secuencia, cambie su punto de interés, perjudicando a las versiones con consideraciones perceptuales.

De entre las causas anteriores la primera se considera la menos probable pues, si se visualizan las secuencias clasificadas, se observa la mejora introducida en el algoritmo. Por su parte, la última de las causas propuestas parece la más acertada: si se observan los datos sobre intervalos de confianza, estos son mayores que las diferencias entre la media de la MOS de las diferentes versiones de las secuencias codificadas. Esto significa que aunque las medias son diferentes, a la vista de las varianzas en las respuestas subjetivas no se puede concluir que las diferencias sean estadísticamente significativas, ni en las pruebas de Julio ni en las de Enero; considerando la MOS sólo un indicativo sin valor estadístico real.

- Una conclusión destacada extraída de estas pruebas subjetivas es la necesidad del clasificador de regiones de interés, además de añadir modificaciones al modo de evaluar las secuencias codificadas. En futuros experimentos y siguiendo la tendencia que parece observarse en algunos artículos sobre consideraciones perceptuales, podría emplearse una cantidad menor de sujetos para visionar cada secuencia, de manera que las diversas opciones de codificación no "entrenen" al individuo con una secuencia en particular. Otros métodos de comparación de doble estímulo podrían emplearse para medir directamente pares de secuencias codificadas

- En cuanto al uso del índice MOVIE, puesto que en las pruebas de Enero existe una cierta correlación con la MOS, podría mantenerse para tomar decisiones sencillas de diseño. Sin embargo, esta medida no puede sustituir a las pruebas subjetivas, pues no es adecuado ajustar el algoritmo de consideraciones perceptuales según los parámetros particulares de una medida particular.

- Por último, otra de las razones por las que se generaron tantas secuencias codificadas diferentes fue para determinar el salto de QP óptimo para cada secuencia. Con respecto a este parámetro no se ha obtenido una conclusión definitiva, aunque parece que para $\Delta QP=$ [redacted] y, en general, para valores

bajos, funciona mejor desde el punto de vista perceptual que para valores altos.

7.3 Mejoras y tareas futuras

El campo de la codificación perceptual de vídeo no ha sido muy explotado por el momento, siendo una vía de investigación interesante en la que trabajar y en la que tiene cabida este estudio. La herramienta de segmentación del movimiento que se propone se considera una aportación a dicha área, y requiere ser perfeccionada aún por medio de tareas como las que se comentan a continuación. Por tanto, para finalizar, se aporta un conjunto de posibles mejoras que realizar para continuar con este trabajo en un futuro; surgen a partir de los dos experimentos realizados sobre las dos versiones del algoritmo de enmascaramiento disponibles hasta el momento.

- Una posible mejora consiste en convertir el clasificador λ_1 en una función de coste que catalogue distintos grados de movimiento, dejando de ser una decisión binaria (mucho/poco movimiento). De esta manera se generaría un mapa de incrementos de QP con cambios más graduales que podrían mejorar la calidad subjetiva. Además, esta mejora no serviría únicamente para introducir las consideraciones perceptuales por movimiento, sino también sería válida para las consideraciones por texturas, por ejemplo, en caso de que interesara implementarlas.

- Los resultados de las pruebas subjetivas han mostrado las carencias del algoritmo de segmentación del movimiento para detectar regiones de mucho y poco interés; de modo que el segundo clasificador ha de desarrollarse para completar la información que proporciona el primero. Hay que señalar que, aunque la opinión subjetiva no puede ser tomada en cuenta por su irrelevancia estadística, es cierto que la detección del punto de interés es la clave para determinar la enmascarabilidad.

- Por otro lado, queda pendiente mejorar los métodos de introducción de distorsión adicional, visto el efecto del incremento de QP y el IZZ, que ya fue descartado en la primera de las pruebas. Se proponen, por tanto, el truncado espectral y el filtrado de *blurring*, que sería adecuado explorar

debido a que cualquier método que asigne recursos de manera desigual tendrá los mismos defectos. En el caso de la primera propuesta, consistiría en obviar coeficientes de alta frecuencia (de la DCT característica de cada MB) que no aportan información relevante de textura; por su parte, el filtrado de *blurring* podría seguir la técnica propuesta en [24].

- Una vez implementado el clasificador de regiones de interés y elegido el método de distorsión apropiado, se debería llevar a cabo un estudio que valide la suposición de partida reflejada en la tabla 3.1 del capítulo 3 referente a la asignación de distorsión.

- Otra de las tareas futuras más importantes es encontrar una variación del experimento subjetivo que permita obtener unos datos más fiables, explorando los experimentos de estímulo sencillo, reduciendo el número de sujetos, o realizando las pruebas de estímulo doble sin referencia, mediante una comparativa de dos secuencias codificadas de manera diferente. Con esta última técnica se garantizaría que en un mismo visionado el observador está evaluando cuál de las dos es mejor y, por tanto, no afecta que se observen las distintas codificaciones de manera distinta.

- Además, queda pendiente, el análisis computacional del algoritmo EMJ con objeto de reducir su coste sin que la calidad del mapa de vectores se vea deteriorada.

En definitiva, es necesario destacar que, a pesar de que los resultados de las pruebas subjetivas no han arrojado los valores objetivo, hay que señalar que los bloques básicos construidos van a permitir construir un sistema adecuado una vez se solventen los problemas mencionados. Es necesario confiar en la aportación del segundo clasificador en el sistema global, que, junto con un estudio de técnicas de introducción de distorsión más apropiado a este escenario, permitirán corroborar la filosofía que rige el sistema objetivo propuesto inicialmente en este estudio.

Capítulo 8

Presupuesto

En este capítulo se realizan los cálculos correspondientes a los costes asociados a la realización de este proyecto. El presupuesto se desglosa en costes de materiales empleados y costes de honorarios de las personas encargadas de llevarlo a cabo. Tras detallar cada uno de los costes se adjunta una tabla resumen que contabiliza los gastos totales acumulados durante el periodo de desarrollo del proyecto.

8.1 Coste del material

El material empleado a lo largo del periodo de trabajo consta del siguiente listado de elementos; se incluyen tanto componentes hardware requeridos como software.

- *PC*. El valor aproximado del equipo informático es de 800 €; como éste puede ser reutilizado tras la realización de este proyecto, su coste puede amortizarse hasta la cantidad de 160 €, considerando un periodo de depreciación de 60 meses.

- *Espacio de trabajo* con las debidas condiciones de luz, calefacción, mantenimiento, más el mobiliario necesario; tiene un coste asociado de unos 900€/mes. Al tratarse de un laboratorio compartido, tendrá un coste individual asociado de 150 €/mes. Debido a que la duración del proyecto ha sido de aproximadamente 12 meses, el coste relativo al espacio de trabajo asciende a 1800 €.

- *Acceso al mini-cluster,* de uso compartido en el departamento, para lanzar las pruebas necesarias relacionadas con la evaluación del proyecto. Su coste estimado es de 100 €.

- *Material de oficina.* Aquí se incluye todo el material desechable utilizado como la impresión de artículos e informes, folios, carpetas, bolígrafos, CDs, etc. Se considera un total de 70 €.

- *Licencias software:*
  - *Sistema operativo Windows XP Professional* de *Microsoft:* licencia de 220 € a amortizar en 4 años, por lo tanto, el coste aplicable a este proyecto es de 55 €.
  - *Matlab R2007b de MathWorks,* utilizado en la implementación del sistema, cuya licencia es de 1950 € (4 años de amortización), supone la cantidad de 487,5 €.
  - *Microsoft Office 2003*, cuya licencia es de 198 €. Considerando una amortización de 4 años, el coste sería de 49,5 €.
  - *Visual Studio 2005 de Microsoft,* utilizado en el desarrollo del sistema también. La licencia de esta herramienta es de 655 €, de modo que el coste amortizándolo en 4 años es de 163,75 €.

- *Conexión ADSL.* La tarifa plana ADSL tiene un coste mensual de 40 €, lo que supone un total de 480 €.

Las siguientes tablas recogen todos los costes relacionados con el material utilizado en el desarrollo del proyecto:

EQUIPOS					
Descripción	**Coste(Euro)**	**% Uso dedicado proyecto**	**Dedicación (meses)**	**Periodo de depreciación**	**Coste imputable**
PC	800 €	100	12	60	160 €
				TOTAL:	160 €

Tabla 8.1. Costes asociados a equipos

OTROS COSTES DIRECTOS DEL PROYECTO			
Descripción	**Empresa**	**Coste imputable**	**Total**
Espacio de trabajo	-	150 €/mes	1800 €
Acceso *mini-cluster*	-	100 €	100 €
Material oficina	-	70 €	70 €
Licencia *Windows XP Professional*	Microsoft	55 €	55 €
Licencia *Matlab R2007b*	MathWorks	487,5 €	487,5 €
Licencia *Microsoft Office* 2003	Microsoft	49,5 €	49,5 €
Licencia Microsoft *Visual Studio* 2005	Microsoft	163,75 €	163,75 €
Conexión ADSL	-	40 €/mes	480 €
		TOTAL:	3.205,75 €

Tabla 8.2. Otros costes directos del proyecto

8.2 Coste de honorarios

Para establecer el presupuesto asociado a los honorarios de las personas a cargo de este proyecto es necesario tener en cuenta en primer lugar, la duración del proyecto y las horas de trabajo realizadas en el mismo. Por tanto, si la duración total es de 12 meses y se establece un horario de trabajo de 6 horas diarias y una dedicación de 5 días semanales, se obtiene un total de 1440 horas laborables.

Para el cálculo completo del coste es necesario tener en cuenta los honorarios de un Ingeniero Técnico de Telecomunicación. Hasta hace poco, la Junta General del Colegio Oficial de Ingenieros Técnicos de Telecomunicación ofrecía una cantidad a modo de ejemplo para los libres ejercientes de la profesión, de forma que pudieran disponer de una referencia de los honorarios. Dicha cantidad definía unos honorarios de 65 €/hora. Hoy en día, el Ministerio de Economía y Hacienda ha remitido a todos los colegios profesionales una nota [37] en la que se indica que no se debe, ni siquiera, publicar un baremo con los

honorarios ya que, éstos son libres y responden al libre acuerdo entre el profesional y el cliente.

Dada la situación, los honorarios deben definirse en función de una serie de factores: costes del ingeniero, desplazamientos, volumen de la actividad, etc. Para este caso concreto es necesario tener en cuenta estos elementos y considerar que, por lo general, un ingeniero técnico que desarrolla trabajos de investigación en la Universidad Carlos III de Madrid dedica unas 1.155 horas (8,8 hombres mes) como máximo al año a dicho fin. Con respecto a los honorarios del director del proyecto, en general, se corresponden con un 7% del coste total del proyecto, lo que supone una dedicación de 0,53 hombres mes o 69,56 horas.

De modo que los gastos personales según los datos descritos son los siguientes:

PERSONAL						
Apellidos y nombre	**N.I.F**	**Categoría**	**Dedicación (*hombre mes)**	**Coste hombre mes**	**Coste (Euro)**	**Firma de conformidad**
Sergio Sanz Rodríguez-Escalona	-	Ingeniero Superior	0,53	4.289,54	2.273,46	-
Ana Belén Mejía Ocaña	-	Ingeniero	8,8	2.694,39	23.710,63	-
		Hombres mes:	9,42	TOTAL:	25.984,09	

*1 Hombre mes = 131,25 horas. Máximo anual de dedicación de 12 hombres mes (1575 horas)
Máximo anual para PDI de la Universidad Carlos III de Madrid de 8,8 hombres mes (1.155 horas)

Tabla 8.3. Coste de honorarios

8.3 Presupuesto total

En definitiva, el coste total del proyecto compuesto por los gastos materiales y de honorarios detallados anteriormente queda reflejado a continuación:

RESUMEN DE COSTES	
Concepto	**Presupuesto Costes Totales**
Personal	25.984
Amortización	160
Subcontratación de tareas	0
Costes de funcionamiento	3.206
Costes indirectos (20%)	5.870
TOTAL:	35.220 €

Tabla 8.4. Presupuesto total

El presupuesto total del proyecto asciende a TREINTA Y CINCO MIL DOSCIENTOS VEINTE EUROS.

Fdo: Ana Belén Mejía Ocaña

Ingeniera Técnica de Telecomunicación, especialidad Sonido e Imagen

Anexo I

Organización DVD adjunto

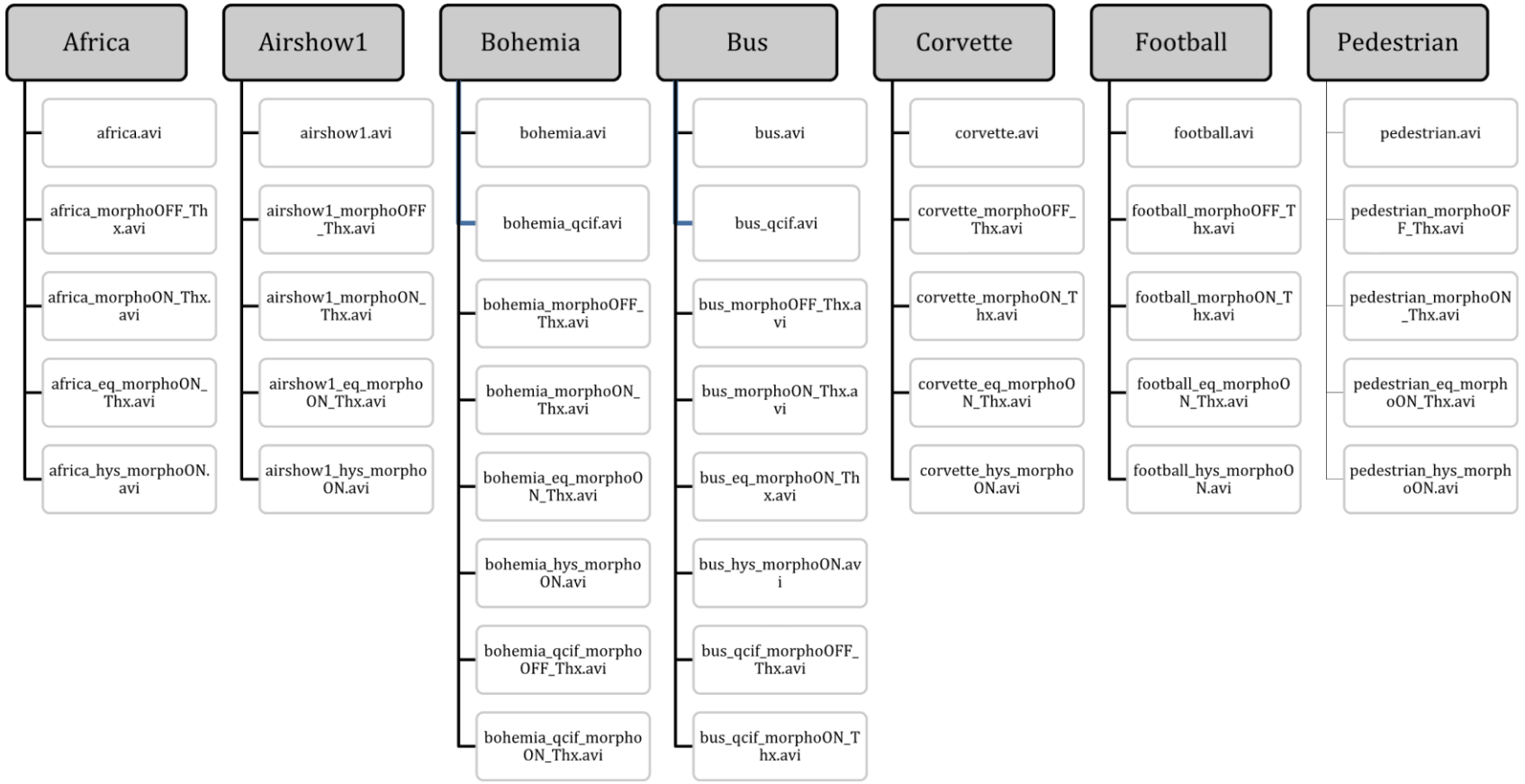
En este anexo se incluye una descripción detallada del contenido del DVD adjunto que contiene secuencias representativas de las pruebas realizadas a lo largo del desarrollo de este proyecto. Están disponibles tanto las secuencias originales como las secuencias características de las diferentes versiones disponibles del algoritmo, para facilitar la comparación de versiones.

El DVD contiene tantos directorios como secuencias de vídeo originales se han utilizado a lo largo del desarrollo del proyecto, cuyos nombres se corresponden con los de los vídeos en cuestión. Adicionalmente, se pone a disposición del lector un reproductor apropiado para la visualización de las secuencias y un archivo "leeme.txt" donde se indica el tamaño de los vídeos, necesario para la reproducción de secuencias SD. Cada carpeta proporciona la secuencia original del vídeo y todas las posibles versiones de dicha secuencia referenciadas. Es necesario indicar que las secuencias de tamaño QCIF a las que se hace referencia a lo largo de este documento se incluyen en el directorio cuyo nombre coincide con el de la misma secuencia, de modo que no se encuentran en un directorio aparte.

La siguiente tabla lista los diferentes nombres que pueden tomar los archivos en función de la versión de la que se trate.

NOMBRE ARCHIVO	***DESCRIPCIÓN***
<vídeo>.avi	- Secuencia original
<vídeo>_qcif.avi	- Secuencia QCIF original
<vídeo>_morphoOFF_Thx.avi	- Versión básica del clasificador λ_1 con un umbral "x".
<vídeo>_morphoON_Thx.avi	- Versión básica del clasificador λ_1 con un umbral "x". - Post-procesado activado.
<video>_eq_morphoON_Thx.avi	- Versión básica del clasificador λ_1 con un umbral "x". - Post-procesado activado. - Pre-procesado activado.
<vídeo>_hys_morphoON.avi	- Versión mejorada del clasificador λ_1. - Post-procesado activado.

Para finalizar, se adjunta un diagrama que refleja de forma esquemática la organización de los archivos en el DVD, indicando qué directorios comprende así como todas las secuencias que contiene cada uno de ellos.


Africa
africa.avi
africa_morphoOFF_Thx.avi
africa_morphoON_Thx.avi
africa_eq_morphoON_Thx.avi
africa_hys_morphoON.avi
Airshow1
airshow1.avi
airshow1_morphoOFF_Thx.avi
airshow1_morphoON_Thx.avi
airshow1_eq_morphoON_Thx.avi
airshow1_hys_morphoON.avi
Bohemia
bohemia.avi
bohemia_qcif.avi
bohemia_morphoOFF_Thx.avi
bohemia_morphoON_Thx.avi
bohemia_eq_morphoON_Thx.avi
bohemia_hys_morphoON.avi
bohemia_qcif_morphoOFF_Thx.avi
bohemia_qcif_morphoON_Thx.avi
Bus
bus.avi
bus_qcif.avi
bus_morphoOFF_Thx.avi
bus_morphoON_Thx.avi
bus_eq_morphoON_Thx.avi
bus_hys_morphoON.avi
bus_qcif_morphoOFF_Thx.avi
bus_qcif_morphoON_Thx.avi
Corvette
corvette.avi
corvette_morphoOFF_Thx.avi
corvette_morphoON_Thx.avi
corvette_eq_morphoON_Thx.avi
corvette_hys_morphoON.avi
Football
football.avi
football_morphoOFF_Thx.avi
football_morphoON_Thx.avi
football_eq_morphoON_Thx.avi
football_hys_morphoON.avi
Pedestrian
pedestrian.avi
pedestrian_morphoOFF_Thx.avi
pedestrian_morphoON_Thx.avi
pedestrian_eq_morphoON_Thx.avi
pedestrian_hys_morphoON.avi

Glosario

ARO:	*Adaptative Rounding Offset*
ASO:	*Arbitrary Slice Ordering*
BS:	*Block Size*
CABAC:	*Context-based Adaptive Binary Arithmetic Coding*
CAVL:	*Context-Based Adaptive Variable Length Coding*
CBR:	*Constant Bit Rate*
CCA:	*Canonical Correlation Analysis*
CFF:	*Critical Flicker/Fusion Frequency*
CIF:	*Common Intermediate Format*
CSF:	*Contrast Spatial Function*
DCR:	*Degradation Category Rating*
DCT:	*Discrete Cosine Transform*
DPCM:	*Differential Pulse Code Modulation*
EE:	*Elemento Estructurante*
EMJ:	*Estimación de Movimiento Jerárquica*
FJND:	*Foveation Just Noticeable Distortion*
FMO:	*Flexible Macroblock Ordering*
GOP:	*Group Of Pictures*
HVS:	*Human Visual System*
IZZ:	*Increase Zero Zone*
ITU:	*International Telecommunication Union*

JME:	*Joint Motion Estimation*
JND:	*Just Noticeable Distortion*
LUT:	*Look-Up Table*
MAD:	*Mean of Absolute Differences*
MAE:	*Mean Absolute Error*
MB:	*Macro Block*
MOS:	*Mean Of Square*
MOVIE:	*MOtion-based Vídeo Integrity Evaluation*
MSE:	*Mean Squared Error*
MVD:	*Motion Vector Difference*
NMSE:	*Normalized Mean Squared Error*
PSF:	*Point Spread Function*
PSNR:	*Peak Signal to Noise Ratio*
QCIF:	*Quarter CIF*
QP:	*Quantization Parameter*
R-Q:	*Rate-Quantization*
ROI:	*Region Of Interest*
RMSE:	*Root Mean Squared Error*
SAD:	*Sum of Absolute Differences*
SD:	*Standard Definition*
SER:	*Signal to Error Ratio*
SNR:	*Signal to Noise Ratio*
SP:	*Simultaneous Presentation*
SR:	*Search Region*
SSD:	*Sum of Squared Differences*
SSIM:	*Structural similarity*
STD:	*Standard Deviation*
VA:	*Visual Attention*
VBR:	*Variable Bit Rate*
VDSI:	*Visual Distortion Sensibility Index*
VFP:	*Visual Fixation Parameter*
VM:	*Vector de movimiento*
VSSIM:	*Vídeo Structural Similarity*

Referencias

[1] C.-W. Tang, C.-H. Chen, Y.-H. Yu and C.-J. Tsai, "Visual Sensitivity Guided Bit Allocation for Video Coding", Proc. IEEE *Trans. Multimedia 8* (Feb. 2006).

[2] H. Yu, F. Pan, Z. Lin and Y. Sun, "A Perceptual Bit Allocation Scheme for H.264", Proc. IEEE ICME 2005.

[3] S. Sengupta, S. K. Gupta and J. M. Hannah, "Perceptually motivated bit-allocation for H.264 encoded video sequences", Proc. IEEE ICIP 2003.

[4] K. Minoo and T. Q. Nguyen, "Perceptual Video Coding with H.264", Proc. IEEE *conference on Signals, Systems and Computers*, 2005.

[5] Y. Sun, I. Ahmad, D. Li and Y.-Q. Zhang, "Region-based Rate Control and Bit Allocation for Wireless Video Transmission", Proc. IEEE *Transactions on Multimedia* Feb. 2006.

[6] S. Moradi, S. Gazor and T. Linder, "A multiple description video coding motivated by human visual perception", Proc. ICASSP 2008.

[7] Laurent Itti, "Automatic Foveation for Video Compression using a Neurobiological Model of Visual Attention", Proc. IEEE *Transactions on Image Processing* Oct. 2004.

[8] J. Chen, J. Zheng and Y. He, "Macroblock-Level Adaptive Frequency Weighting for Perceptual Video Coding", Proc. IEEE *Transactions on Consumer Electronics* April 2007.

[9] S. Xu, Li Yu and G. Zhu, "A perceptual coding method based on the luma sensivity model", Proc. IEEE ISCAS 2007.

[10] Henry H. Y. Tong and Anastasios N. Venetsanopoulos, “A perceptual model for JPEG applications based on blocks classifications, texture masking, and luminance masking”, Proc. IEEE ICIP 1998.

[11] Xin Jin, Satoshi Goto and K. N. Ngan, “Optical flow based DC surface compensation for artifacts reduction”, Proc. PCS2009.

[12] Z. Chen and C. Guillemot, “Perception-oriented video coding based on foveated JND model”, Proc. PCS2009.

[13] U. Engelke, H.-J. Zepernick and A. Maeder, “Visual attention modeling: Region-Of-Interest versus Fixation Patterns”, Proc. PCS2009.

[14] A. Bhat, I. Richardson and S. Kannangara, “A novel perceptual quality metric for video compression”, Proc. PCS2009.

[15] S. Li, S. Chen, J. Wang and Lu Yu, “Second order prediction on H.264/AVC”, Proc. PCS2009.

[16] K. Panusopone, Jae H. Kim, and Limin Wang, “Joint block motion estimation in H.264/AVC”, Proc. PCS2009.

[17] M. Bosch, F. Zhu and Edward J. Delp, “An overview of texture and motion based coding at Purdue University”, Proc. PCS2009.

[18] H. R. Wu and K. R. Rao, “Digital Video Image Quality and Perceptual Coding”, 2006 by Taylor & Francis Group, LLC.

[19] Andrew T. Duchowski, “Eye Tracking Methodology”, Theory and Practive, Second Edition, 2007 Ed. Springer.

[20] Iain E. G. Richardson, “H.264 and MPEG-4 Video Compression”, Ed. Wiley, 2003.

[21] Sandra Torrades and Pol Pérez-Sust, “Sistema visual, la percepción del mundo que nos rodea”, Proc. OFFARM, vol. 27 núm. 6, Jun. 2008.

[22] J. Han, M. Li, H. Zhang and Lei Guo, “Automatic Attention Object Extraction from Images”, Proc. IEEE, ICIP 2003.

[23] Francesc Tarrés Ruiz “Sistemas audiovisuales, Televisión analógica y digital”, Ediciones UPC, 2000.

[24] J.-S. Lee, F. De Simone and T. Ebrahimi, “Video Coding based on Audio-Visual Attention”, Proc. ICME 2009.

[25] Y. Huang, K. Palaniappan, X. Zhuang, and J. E. Cavanaugh, “Optic flow field segmentation and motion estimation using a robust genetic partitioning

algorithm", IEEE *Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, Nº 12, December 1995.

[26] C. R. Blanco, F. Jureguizar, L. Salgado, and N. García, "Target detection through robust motion segmentation and tracking restrictions in aerial flir images", IEEE ICIP 2007.

[27] Wilhelm Burger, Mark James Burge, "Digital image processing: an algorithmic introduction using Java", Ed. Springer, 2008.

[28] Q. Huynh-Thu, and M. Ghanbari, "Asymmetrical temporal masking near video scene change", Proc. IEEE, ICIP 2008.

[29] "ITU-T P.910 (04/2008): Subjective video quality assessment methods for multimedia applications".

[30] "Recommendation ITU-R BT.500-11: Methodology for the subjective assessment of the quality of television pictures".

[31] K. Seshadrinathan and A. C. Bovik, "Motion Tuned Spatio-temporal Quality Assessment of Natural Videos", vol. 19, no. 2, pp. 335-350, *IEEE Transactions on Image Processing*, Feb. 2010.

[32] WATSON, A.B. "The cortex transform: rapid computation of simulated neural images", *Computer Vision, Graphics, and Image Processing*, vol. 39, pp. 311-327, 1987.

[33] Carlos Esteban Baz Hormigos, Manuel de Frutos López, "Medidas de calidad subjetiva en secuencias de vídeo", PFC, Sep. 2009.

[34] K. Seshadrinathan and A. C. Bovik, "Motion-based Perceptual Quality Assessment of Video", SPIE *Conference on Human Vision and Electronic Imaging*, 2009

[35] Q. Xu, X. Lu, Y. Liu and C. Gomila, "A Fine Rate Control Algorithm with Adaptive Rounding Offsets (ARO)", *IEEE Transactions on circuits and Systems for Vídeo Technology*, Oct. 2009.

[36] Chen, Z. & Ngan, K. N., "Recent advances in rate control for video coding", *Image Commun*., *Elsevier Science Inc*., 2007, 22, 19-38.

[37] COITT. Último acceso: 4/09/2009
http://www.coitt.es/res/libredocs/Honorarios.pdf